Developing a video-based method to compare and adjust examiner effects in fully nested OSCEs

Peter Yeates,[1,2] Natalie Cope,[1] Ashley Hawarden,[3] Hannah Bradshaw,[3] Gareth McCray[4] & Matt Homer[5]

BACKGROUND Although averaging across multiple examiners' judgements reduces unwanted overall score variability in objective structured clinical examinations (OSCE), designs involving several parallel circuits of the OSCE require that different examiner cohorts collectively judge performances to the same standard in order to avoid bias. Prior research suggests the potential for important examiner-cohort effects in distributed or national examinations that could compromise fairness or patient safety, but despite their importance, these effects are rarely investigated because fully nested assessment designs make them very difficult to study. We describe initial use of a new method to measure and adjust for examiner-cohort effects on students' scores.

METHODS We developed video-based examiner score comparison and adjustment (VESCA): volunteer students were filmed 'live' on 10 out of 12 OSCE stations. Following the examination, examiners additionally scored station-specific common-comparator videos, producing partial crossing between examiner cohorts. Many-facet Rasch modelling and linear mixed modelling were used to estimate and adjust for examiner-cohort effects on students' scores.

RESULTS After accounting for students' ability, examiner cohorts differed substantially in their stringency or leniency (maximal global score difference of 0.47 out of 7.0 [Cohen's d = 0.96]; maximal total percentage score difference of 5.7% [Cohen's d = 1.06] for the same student ability by different examiner cohorts). Corresponding adjustment of students' global and total percentage scores altered the theoretical classification of 6.0% of students for both measures (either pass to fail or fail to pass), whereas 8.6–9.5% students' scores were altered by at least 0.5 standard deviations of student ability.

CONCLUSIONS Despite typical reliability, the examiner cohort that students encountered had a potentially important influence on their score, emphasising the need for adequate sampling and examiner training. Development and validation of VESCA may offer a means to measure and adjust for potential systematic differences in scoring patterns that could exist between locations in distributed or national OSCE examinations, thereby ensuring equivalence and fairness.

Medical Education 2019: 53: 250–263
doi: 10.1111/medu.13783

[1]Medical School Education Research Group (MERG), Keele University School of Medicine, Keele, UK
[2]Department of Acute Medicine, Fairfield General Hospital, Pennine Acute Hospitals NHS Trust, Bury, UK
[3]Royal Stoke Hospital, University Hospital of North Midlands NHS Trust, Stoke on Trent, UK
[4]Institute for Primary Care and Health Sciences, Keele University, Keele, UK
[5]School of Education, University of Leeds, Leeds, UK

Correspondence: Peter Yeates, School of Medicine, David Weatherall Building, Keele University, Keele, Staffordshire ST5 5BG, UK.
Tel: 00 44 1782 733930; E-mail: p.yeates@keele.ac.uk

BACKGROUND

Fairness in assessments is a vital part of the educational contract that students have with their institutions[1] and standardisation helps to reassure the public that all graduates have met predefined assessment criteria.[2] For these reasons, despite advances in programmatic assessment,[3] entrustability frameworks,[4] narrative judgements[5] and competency-based medical education,[6] summative assessments for graduation or licensing purposes typically continue to use single, high-stakes assessments that strive toward equivalent assessment under strict but fair conditions. This study describes an innovative approach to understanding, and seeking to enhance, a rarely considered aspect of fairness in such examinations.

Within high-stakes summative assessments, learners' clinical skills are usually assessed by objective structured clinical examinations (OSCEs)[7] or closely related variations such as standardised patient assessments.[8] A considerable body of literature has examined the validity of these assessments from a predominantly psychometric perspective,[9] although sociocultural critiques have also been made.[10] Several factors have established influences on the reliability of OSCEs: the number of stations and testing time;[11] the number of examiners per station;[8] content specificity effects arising from station tasks;[12] and the format of scoring responses.[13]

Examiner variability is often a substantial source of construct-irrelevant variance in OSCEs.[9,14] Training examiners is strongly recommended and some empirical findings support its benefits.[15,16] In the original conceptions of OSCEs, all students were intended to meet all examiners,[17] and as such examiner variability was unlikely to advantage or disadvantage particular students unless examiners showed idiosyncratic behaviour towards subsets of students.[16] Owing to student numbers, most contemporary OSCEs are conducted either across multiple simultaneous parallel circuits in the same location or at different geographical locations. Examiners tend to examine in a single circuit or location for several cycles of students, and as a result, each student is examined by only a subset of examiners (or by one 'examiner cohort'[18]). It is consequently critical to the fairness of OSCEs that each different cohort of examiners (in different parallel circuits or different locations) collectively judge performances to the same standard of judgement to ensure that students are not systematically either advantaged or disadvantaged by the circuit or location in which they perform.

Comparatively few studies have examined the influence of different circuits on OSCE examinations. Tamblyn et al.[19] experimentally compared ratings of examiners from two different sites by asking them to rate a small subset of videos that had been obtained in an OSCE. They showed that although inter-examiner agreement was identical within each site, there was a systematic difference of 6.7% between the two sites. Extrapolating their findings to the real OSCE would have significantly influenced pass and fail rates. Early studies by de Champlain et al.[20] and Reznick et al.[21] did not demonstrate any influence of assessment site on scores, whereas more recently, Floreck and de Champlain[22] examined differences across 21 sites in the USA and found that examination site explained between 3.0% and 11.6% of score variance. Sebok et al.[23] analysed aggregated data to compare examiner effects across sites. They found that site differences variably explained between 1.5% and 17.1% of score variability. Yeates and Sebok-Syer[18] specifically addressed whether parallel examiner cohorts across different sites in the same medical school showed different standards of judgement. Their provisional results suggested that scores by different examiner cohorts differed by up to 4.4% of the assessment scale.

In summary, heterogenous findings have been reported across different studies for the influence of different sites or different groups of examiners on OSCE scores. One difficulty with most of these studies is that it is unclear whether the differences that were observed represent differences in judgements by different examiner cohorts (i.e. error) or genuine differences in the abilities of students in each location (true score variance). Studying these effects robustly is often difficult or impossible within standard OSCE designs because students are usually fully nested within cohorts of examiners, with no crossover between groups of examiners and groups of students. Although estimation of the influence of different circuits on an OSCE has previously been attempted,[24] direct comparisons are usually impossible because student ability and the standard of examiners' judgement are confounded.

Despite this difficulty, addressing these differences is educationally highly important. These studies suggest the potential for differences between examiner cohorts or locations that could importantly impact fairness within assessments. Moreover, there could be very important implications for patient safety if licensing examinations operate to different standards of judgements in different geographical locations. Despite this, in the common scenario that examiners are fully nested within subsets of students, no established method currently exists to robustly measure the influence of different examiner cohorts or geographical locations within a single OSCE examination, which does not rely on assumptions about the distribution of students' abilities.

The aim of this study was to describe the development of a novel combination of practical steps, paired with established statistical analytical methods, to produce a method that may be capable of jointly addressing the difficulties posed by fully nested OSCE designs, without the need for such assumptions. Using this method, we sought to determine the following.

1 How the standard of judgement compares between different fully nested examiner cohorts in parallel circuits of an OSCE examination?
2 What influence adjusting for any such differences might have on students' scores?

In addressing these aims, we sought to provide data and experience that will enable further development of this method.

METHODS

Overview

We used a novel combination of processes that we called 'video-based examiner score comparison and adjustment' (VESCA). This involved three procedures in sequence: (i) a subset of students were filmed 'live' when performing the majority of stations in their real OSCE; (ii) examiners from each of the separate parallel circuits of the examination scored station-specific common-comparator videos of students' performances in the OSCE, and (iii) statistical analyses used the partial crossing created by examiners' scores for the common-comparator (video) performances to estimate the influence of each different examiner cohort on students' scores and to adjust accordingly. Although several examinations have previously used statistical adjustment of students' scores, as far as we can establish, the scoring of station-specific common-comparator videos by examiners as a means to overcome a fully nested design makes these processes novel.

Assessment format

The study was carried out within Year-3 OSCE at Keele University's School of Medicine's 5-year undergraduate medical degree. Students perform one OSCE per year; passing the OSCE is required for progression, although one re-sit is allowed. The 12-station OSCE comprised consultation skills, physical examination and procedural skills. Simulated patients were used in most stations, with real patients involved in two out of the 12 stations. All examiners were experienced clinicians and had undertaken OSCE examiner training (including video-based benchmarking), received detailed station information in advance of the OSCE and attended a pre-OSCE briefing and standardisation exercise.

Examiners allocated scores using Keele's domain-based marking scheme, known as GeCos.[25] Subscales are scored from 1 to 4 ('Must improve in this category' to 'Very good in this category') and summed. Each station had between five and six subscales. Additionally, examiners scored a 7-point global score ranging from 1 (incompetent) to 7 (excellent), which was added to the sum of the domain-scale scores for each question, giving a total score for each question out of either 27 points (where there were five subscales) or 31 points (where there were six subscales). Cut-off scores for each station were calculated using borderline regression,[26] derived from a further 5-point standard setting scale, which was not included in this study as standard setting was not the focus of our inquiry. Examiners scored students' performance and recorded verbal feedback on electronic tablet devices using Keele's electronic OSCE feedback platform.[27]

Owing to student numbers, the OSCE was conducted in four simultaneous parallel circuits (referred to as red, blue, green and orange lanes), each relying on different groups of examiners to deliver ostensibly the same OSCE. Station scenarios were the same across the four lanes. The OSCE was split over 3 days, with all students examined on the same four stations on each day, and attending on

all three successive days. Students rotated through the OSCE in groups of five of six students, termed 'cycles', with three cycles of students in the morning and two cycles of students in the afternoon. Students were allocated to circuits sequentially based on their student numbers, which are generated in an essentially random process and are not expected to produce any systematic groupings. Students were examined in the same cycle in the same lane on each day. The layout of the parallel circuits and cycles of the OSCE are illustrated in Appendix S1.

The majority of examiners judged the OSCE for half a day; a minority remained all day or returned for a second half-day on a different day (therefore examining a different station). Nonetheless, there were eight different cohorts of examiners (the morning and afternoon cohorts for each of the four parallel circuits), with only limited recurrence of examiners between cohorts.

VESCA phase 1 procedures: filming

Following recruitment e-mails to the entire year group, five students volunteered to be filmed during their OSCE and provided written consent. These students were allocated to the red lane in the first cycle of the morning of each day and were unobtrusively filmed (using ceiling-mounted video cameras and hanging microphones) on 10 out of the 12 OSCE stations. The remaining two stations were excluded as they featured real patients and were outside the remit of the ethical approval. Two of the five available video performances were selected for each station on pragmatic grounds (generally the first two students to rotate through that station) to be shown to examiners in phase 2.

VESCA phase 2 procedures: video scoring

Videos were segmented by IT staff to present the portion of time from students entering the station to them leaving, with neither scores nor examiners' audio feedback to students included. After both morning and afternoon sessions, the examiners from all four parallel lanes were invited to score comparison videos. Collectively, participating examiners from all examiner cohorts judged the same comparator videos, but each examiner only judged the two selected videos that were specific to the station they had examined. Despite having already scored the video performances 'live', examiners from the filmed circuit also scored the comparator videos so that their live and video scores could be compared. Examiners watched the performances in the same order on tablet computers, using earphones. They scored the performances and provided written feedback on paper versions of the electronic OSCE mark sheet. Filming and video scoring procedures were repeated on all 3 days of the OSCE, and as a result each of the eight examiner cohorts scored common video performances on 10 of the 12 stations.

VESCA phase 3: analysis

Scores were collated from all students' live performances, along with all examiners' scores for comparator videos. The total possible score on each station varied (either 27 or 31 marks), which meant that they contributed different weights to students' total scores for the OSCE. As we judged that this could bias estimates of station difficulty, we opted to remove this weighting by converting total scores for each station to percentages. The study outcome measures comprised (i) the global scores for each student on each station and (ii) the percentage scores for each student on each station.

A Bland-Altman plot[28] was used to investigate whether any systematic bias existed between live and video scoring of performances, by comparing the subset of scores given by examiners who had provided both live and video-based scores on the same student performances.

We chose many-facet Rasch modelling (MFRM)[29] to estimate the relative influence of examiner cohorts on students' global scores because global scores were ordinal with a small number of response categories. We modelled facets of student, station and examiner cohort. The analysis was run in FACETS by Winstep,[30] which produces estimates for examiner-cohort stringency, station difficulty and student ability. It also routinely provides model-adjusted 'fair score' estimates (i.e. controls for examiner-cohort stringency) as well as parameter estimates and fit statistics. We adopted the fit parameters recommended by Linacre;[31] that is, that infit and outfit mean square values between 0.5 and 1.5 are considered useful for productive measurement whereas infit and outfit *Z*-score values outside $\pm$ 2.0 indicate that the corresponding mean square values are statistically significantly different from 1.0. As a result, these measures indicate different features: the mean square values indicate the extent to which an item within a facet fits or misfits the pattern expected by the model, whereas the z-standardised values indicate the likelihood of

any variations having occurred by chance. No stations, students or examiner cohorts were removed from the analysis on the basis of fit (or any other) criteria. In support of the assumption of unidimensionality, we examined station global score to total global score Spearman's correlations for each station to determine whether stations contributed similarly to the overall global score.

We chose linear mixed modelling (LMM) to estimate the relative influence of examiner cohorts on students' total percentage scores. We chose LMM rather than MFRM because Rasch modelling is more appropriate for the analysis of binary and ordinal categorical data with relatively few response categories. The LMM fitted a linear mixed model to the entire dataset, using the following model:

$$Y_{ij} = \beta_0 + \beta_1 Station_{ij} + \beta_2 Cohort_{ij} + \alpha_i + \varepsilon_{ij}$$

where Y_{ij} is a dependent variable representing total percentage score on observation j for student i, β_0 is the model intercept, β_1 is the coefficient representing the effect of station on the dependent variable, β_2 is the coefficient representing the effect of examiner cohort on the dependent variable, α_i is a random effect representing the underlying student ability (relative to the sample of students) and ε_{ij} is an overall error term. These analyses were performed in R[32] using LME4.[33] R^2 values were extracted using the Nakagawa and Schielzeth[34] method in the r package MuMIn.[35] As a follow-up to examine the proportions of variance explained by the explanatory variables, we calculated the relative importance of station, cohort and student 'ability' using the 'relaimpo'[36] package on a linear model. Although other methods of score adjustment could have been selected (e.g. mean equating[37]), these would have relied on an assumption that student ability was equal across examiner cohorts. We chose our selected analytical methods as they enabled us to control for both examiner bias and student ability concurrently.

Percentage total scores were derived from multiple ordinal scales and might therefore be considered non-interval. Prior research, however, has demonstrated that data that are summed or averaged from multiple Likert items behave similarly to interval data[38] and parametric tests are robust for their analysis.[39,40] Moreover, summed or averaged data from multiple ordinal responses are commonly treated as intervals within assessment procedures in many institutions globally.

Next, we examined the distributions of the differences between students' raw and adjusted scores and changes in classification around a cut-score. Data from the OSCE were supplied by the institution on the condition that we would not use the actual cut-score from the OSCE to model alterations to pass or fail decisions, in case this produced concerns among students. Instead, for the purposes of understanding how the VESCA methodology operates, we examined its influence around a similar cut-score. We derived this cut-score using the borderline regression method,[26] but interpolated from a different point on the standard setting scale (i.e. the *x*-axis), which was within 0.5 scale points of the interpolation point used to set the actual standard. Similarly, we examined the influence around point 4 out of 7 on the global score, which denotes 'satisfactory' performance. We then compared the proportions of students who passed, failed or were reclassified (pass to fail, or fail to pass) between the raw and adjusted scores for both the total percentage scores and global scores.

RESULTS

Completion rates and descriptive data

A total of 116 students were examined. All five volunteer students completed filming and consented to use of their videos. A total of 67 unique examiners observed students within the 10 included stations; 13 examiners observed on two occasions. A total of 49 examiners agreed to take part in Stage 2 (including all 13 examiners who examined twice), giving a response rate of 73.1%. Examiner participation rates differed by examiner cohort (cohort 1, 80.0%; cohort 2, 80.0%; cohort 3, 60.0%; cohort 4, 70.0%; cohort 5, 50.0%; cohort 6, 70.0%; cohort 7, 100%; cohort 8, 100%) and correspondingly estimates of the influence of each examiner cohort were based on scores from between 10 and 20 comparator videos. Scores for video performances comprised 7.9% of the total dataset.

The OSCE had a Cronbach's alpha across stations of 0.62. Examiners used the full range of the global scale (ratings 1–7; median, 5.0; interquartile range [IQR], 2). Students' average global scores ranged from 3.6 to 5.8, with a mean of 4.7 and a standard deviation of 0.47. Percentage total scores ranged

from 32.3% to 100.0% for individual students on individual stations. By contrast, the scores given by individual examiners to individual video performances ranged from 48.1% to 100.0%, with a mean value of 75.5%. Students average percentage total scores (i.e. the percentage of total they scored for the whole OSCE) had a mean of 75.9% and a standard deviation of 5.4%. Collectively, these data suggest that the video performances showed a range of student ability that was broadly similar to the rest of the OSCE. Unadjusted station difficulty ranged from easiest at a mean total percentage score of 79.1% to hardest with a mean total percentage score of 70.9%.

Comparison between live and video scoring

A total of 20 student performances were scored both live and via video by the same examiners, with a delay between these ratings of approximately 2.5 hours. Bland–Altman plots (Fig. 1) demonstrated mean differences for both measures that did not statistically significantly differ from zero. Furthermore, there seemed to be no obvious bias regarding the size or direction of the difference with change in underlying value of the measure (Fig. 1).

Influence of examiner cohorts on overall global scores (MFRM)

Fit statistics from the many-facet Rasch model indicated good fit between the data and the model. All examiner cohort facets had infit and outfit mean square values between 0.5 and 1.5, and infit and outfit z-standardised values of ±2.0, indicating values that were productive for measurement. Standard errors for each examiner cohort were similar, with a median of 0.065 logits and a range from 0.06 to 0.08 logits. This indicates that error variability in examiner cohorts was similar (Table 1).

All stations had infit and outfit mean square values between 0.5 and 1.5. A total of 8 out of 12 stations had infit and outfit z-standardised values of ±2.0. A total of 4 stations showed infit or outfit z-standardised values outside of this range: station 8 (infit, 3.2; outfit, 3.2); station 6 (infit, 2.6; outfit, 2.5); station 2 (infit, −2.1; outfit, −2.0); station 10 (infit, −2.3; outfit, −2.3). Notably, although these z-standardised values indicate that the corresponding infit and outfit mean square values were statistically significantly different to 1.0, the fact that their mean square values were within the range 0.5–1.5 indicates that the magnitude of these deviations was small, and that correspondingly the fit of these stations was still productive for measurement. Station score to total score Spearman's correlations were similar for all stations, with a median of rho = 0.40 and a range of rho = 0.31 to 0.52.

The majority (78.4%) of students' values showed good fit with both mean square values and z-standardised values within the productive measurement range. Moderate underfit was shown by 9.4% of students, with mean square values >1.5-≤2.0; however, none of the z-standardised values for these students was outside of ±2.0, indicating that these differences were not statistically significant. Three students (2.6%) showed more pronounced underfit, with mean square values >2.0. All of these students also showed z-standardised values outside of ±2.0, indicating this underfit was statistically significant. Overfit was shown by 9.4% of students, with mean square values <0.5. Of these, only four students (3.4%) had z-standardised values outside of ±2.0.

Examiner cohorts differed in their stringency or leniency, with a model-adjusted fair score of 4.53 out of 7.0 units of the assessment's global score for examiner-cohort 3 and a model-adjusted fair score of 5.00 out of 7.0 units of the assessment's global score for examiner-cohort 6, a difference of 0.47 out of 7.0 (6.9% of the global scale). These results indicate that a student of a given ability (at the middle of ability distribution) examined by examiner-cohort 3 received a score 0.47 global-scale points lower than a student of the same ability examined by examiner-cohort 6. As the standard deviation of students' ability on the model-adjusted global score was 0.49 units of the global scale, this represents a Cohen's *d* effect size of 0.96 for the maximal difference between examiner cohorts. The Rasch separation reliability was 0.79, indicating that these facets could be reliably separated. These data are shown in relation to the influence of the other facets in the Wright Map in Fig. 2.

Influence of examiner cohorts on total percentage scores (LMM)

The marginal R^2 for the model, expressing the amount of variance the fixed effects (i.e. station and cohort) explained, was 0.09. The conditional R^2, expressing the amount of variance the fixed and random effects (i.e. student 'ability') jointly explain, was 0.17. When broken down, station explained 26.1%, examiner cohort 5.7% and student ability 68.1% of the total score variance. As with the

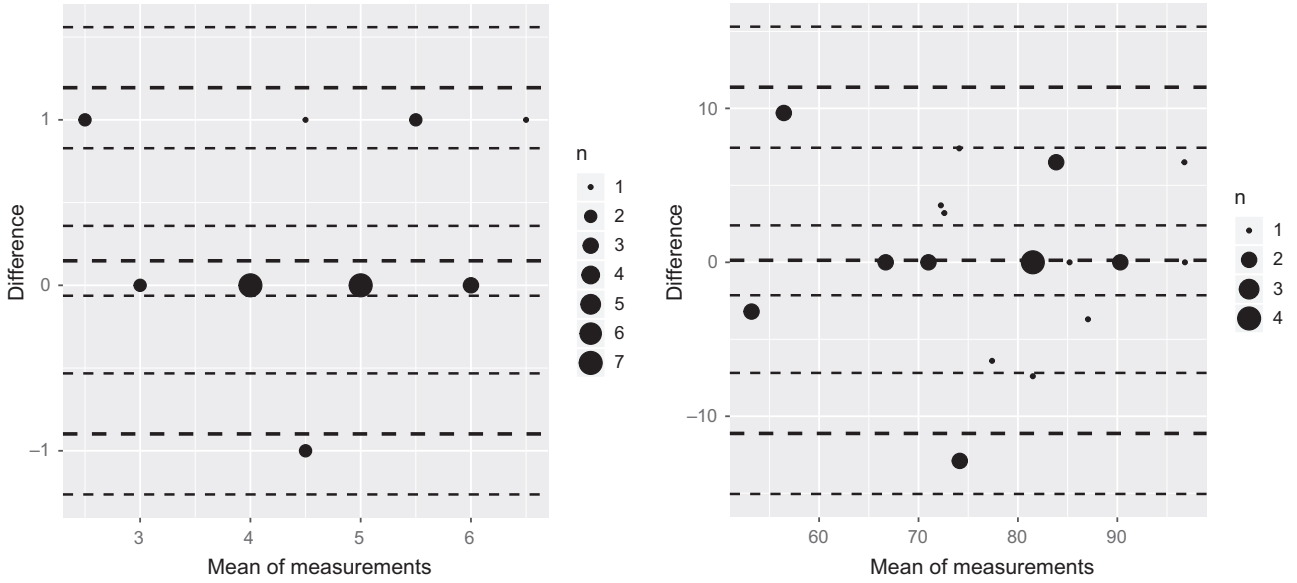

Figure 1 Bland–Altman plots of global score (left) and total percentage score (right); positive difference indicates live scored higher than video. Bold dotted lines represent the mean difference and 95% confidence intervals [CIs] (limits of agreement) of the differences, and the light dotted lines represent the 95% CIs for these values

Table 1 Standard error values for each examiner cohort derived from the many-facet Rasch modelling (in logits) and the linear mixed modelling (in percentage points)

Examiner cohort	MRFM Standard error (logits)	LMM Standard error (percentage score points)
1	0.06	1.44
2	0.08	1.70
3	0.06	1.53
4	0.08	1.70
5	0.06	1.54
6	0.08	1.71
7	0.06	1.48
8	0.07	1.64

LMM = linear mixed modelling; MFRM = Many-facet Rasch modelling

MFRM, standard errors were similar for examiner cohort, with a median of 1.59% and a range of 1.44% to 1.71%, again indicating that error variance was similar across examiner cohorts (Table 1).

The model showed that five of the remaining seven examiner cohorts were statistically different ($p< 0.05$) from the adjusted estimate for the lowest scoring examiner cohort (examiner-cohort 4), indicating differences between these examiner cohorts in their scoring tendencies. Score adjustments for the examiner cohorts (i.e. the scores the different examiner cohorts would give to the same student performance) ranged from −3.2% for examiner-cohort 4 to +2.5% for examiner-cohort 5 (relative to the mean cohort adjustment), a difference of 5.7%. Given that the standard deviation for the adjusted values for student's ability on this measure was 5.4%, this represents a Cohen's *d* effect size of 1.06. These data are shown in relation to the influence of the other facets in Fig. 3.

Effect of adjusting for influence of examiner cohorts on students' scores

The distribution of differences between the raw global scores and their corresponding model-adjusted global scores had a root mean squared error (RMSE) of 0.14 and a standard deviation of 0.16. The greatest increase in global scores was from 3.83 raw to 4.05 adjusted (difference of 0.22 units of global score [3.1%]), whereas the greatest decrease in scores was from 4.08 raw to 3.78 adjusted (difference of 0.30 units of global score [4.3%]). The percentage of students whose score changed by at least ±0.24 (equivalent to a Cohen's *d* of 0.5) was 8.6%.

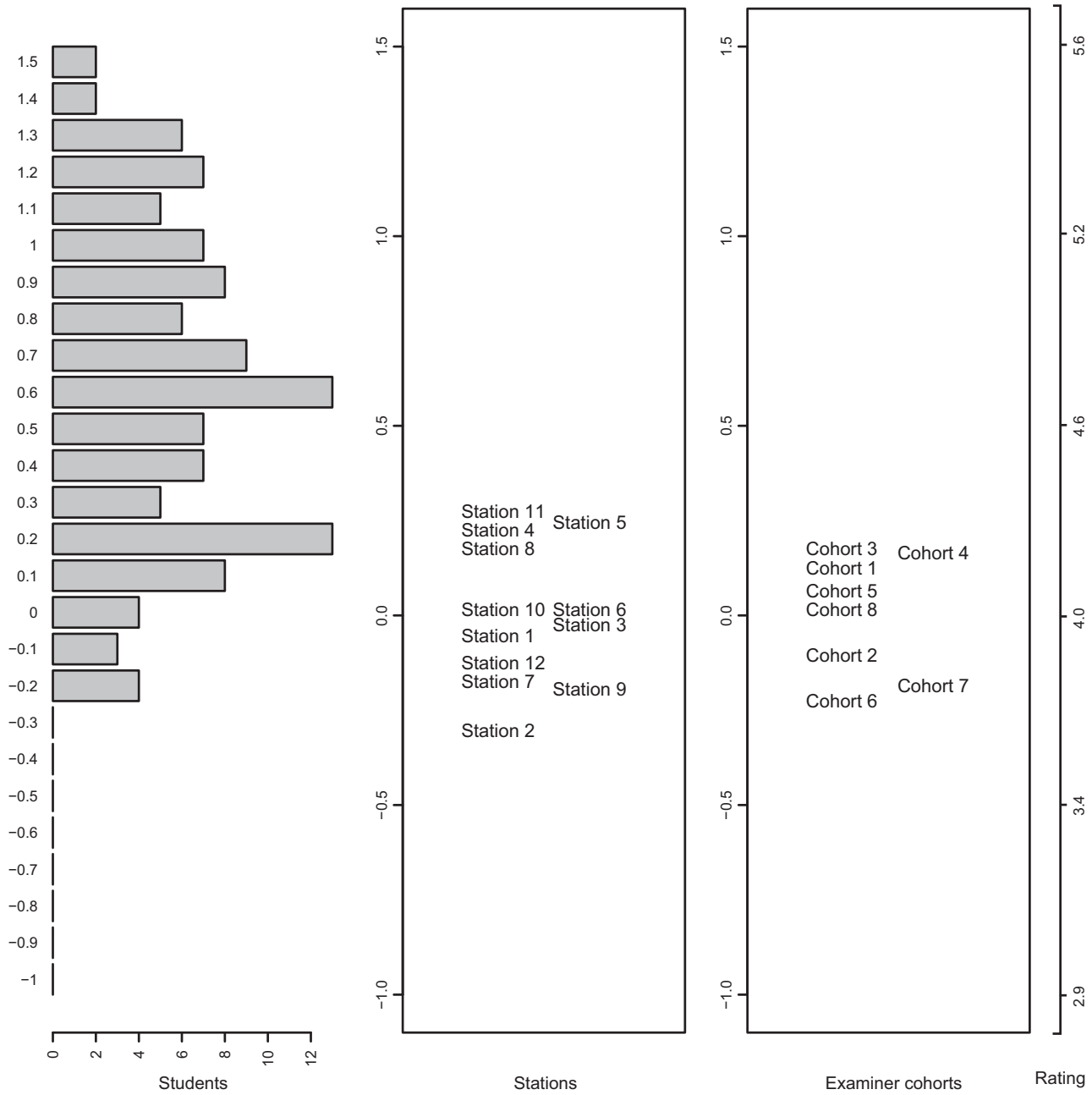

Figure 2 Wright map showing relative influence of students, stations and examiner cohorts on overall global scores

Comparing students' model-adjusted global scores with a cut-score of 4.0 out of 7.0 produced a change in classification for 6.0% of students. Of these, three students (2.6%) passed who would otherwise have failed, whereas four students (3.4%) failed who would otherwise have passed. Five students (4.3%) failed by both methods and the remainder (89.7%) passed by both methods. These data are shown in Fig. 4. Of the students who changed classification, only one showed very mild underfit (MnSq = 1.52), which was not statistically significant (z-standardised = 1.3). The remainder of the reclassified students showed good fit to the model.

The distribution of differences between the total percentage raw scores and their corresponding model-adjusted fair scores had an RMSE of 1.69% and a standard deviation of 1.96%. The greatest increase in scores was from 66.58 to 69.14 (difference of 2.56%), whereas the greatest decrease in scores was from 84.92 to 81.74 (difference of −3.18%). The percentage of students whose score

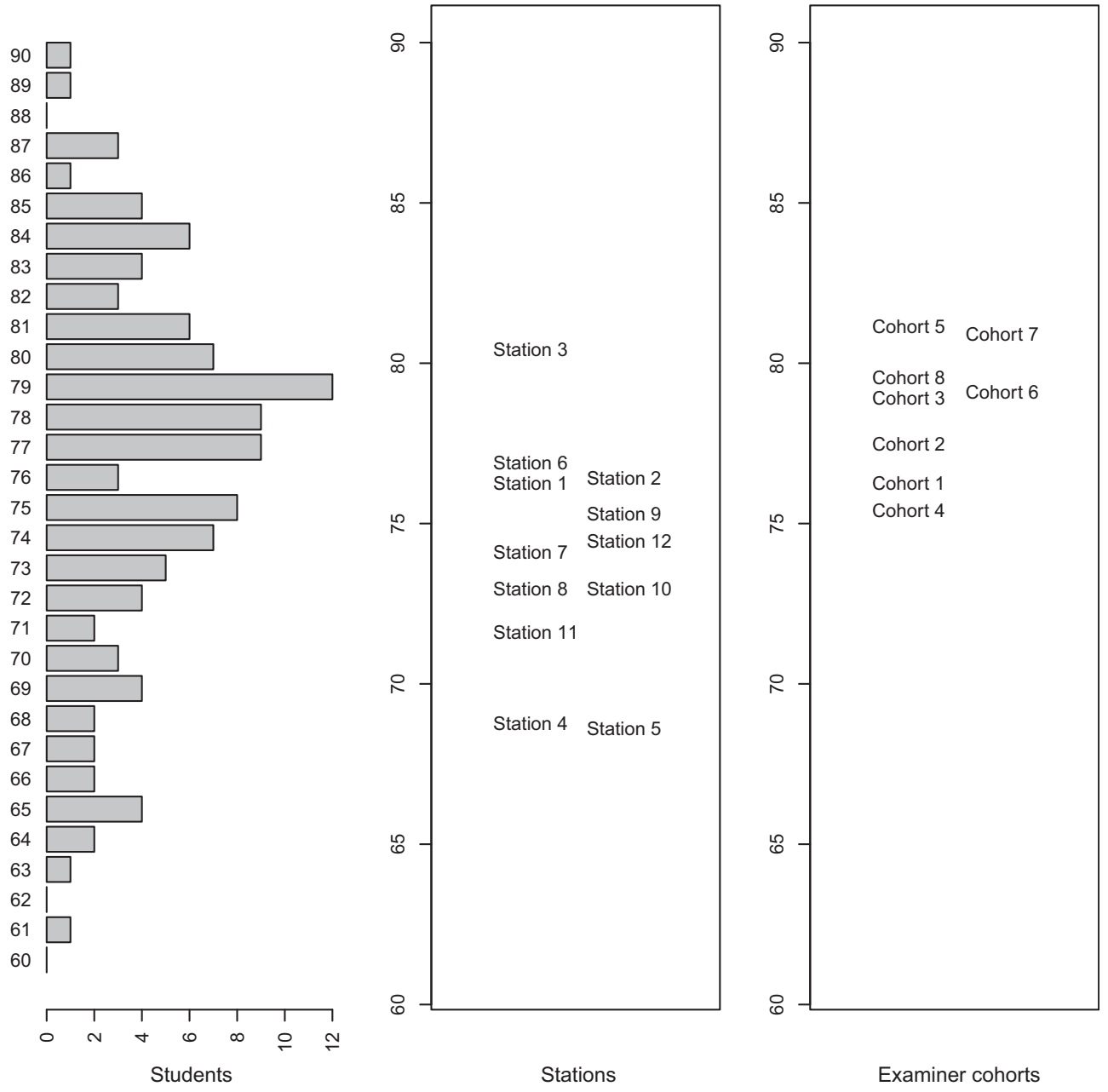

Figure 3 Diagram showing the relative influence of students, stations and examiner cohorts on total percentage scores

changed by at least 2.7% (equivalent to a Cohen's *d* of 0.5) was 9.5%.

Borderline regression produced an artificial cut-score of 67.4%. Comparing students' model-adjusted total percentage scores with this cut-score produced a change in classification for 6.0% of students. Of these, one student (0.8%) passed who would otherwise have failed, whereas six students (5.2%) failed who would otherwise have passed. Six students (5.2%) failed by both methods and the remaining 103 students (88.8%) passed by both methods. These data are shown graphically in Fig. 4.

DISCUSSION

Summary of results

In this study, we have described the preliminary use of VESCA as a novel intervention to measure and adjust for the influence of different examiner

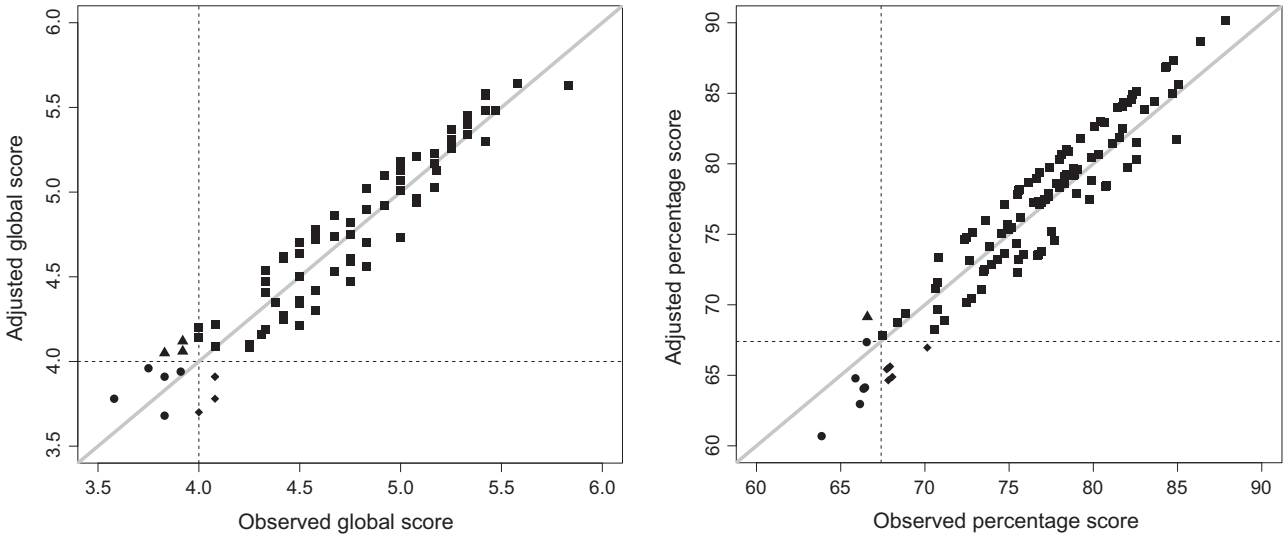

Figure 4 Plot of raw versus model-adjusted overall global scores for individual students for global and total percentage scores. Note: students indicated by triangles (▲) changed from fail to pass when scores were adjusted; students indicated by diamonds (♦) changed from pass to fail when scores were adjusted; students indicated by squares (■) passed under both conditions; and students indicated by circles (●) failed under both conditions

cohorts within a single, fully nested OSCE examination. Examiners showed no systematic differences in their scoring of live and video-based performances, and use of the three-stage procedure of: (i) videoing students' live OSCE performances; (ii) asking examiners to score station-specific common-comparator videos, and (iii) comparing and adjusting for the influence of examiner cohorts proved feasible. Examiner cohorts differed in their leniency or stringency, accounting for differences of up to 6.9% (Cohen's d = 0.96) in global scores and 5.7% (Cohen's d of 1.06) in total percentage scores. Notably, examiner cohorts' rank ordering of leniency and stringency differed across the two measures. Use of model-adjusted scores changed students' classifications around our artificial (but similar to the real) cut-scores from fail to pass for 3.4% and 0.8% of students with global scores and total percentage scores, respectively, whereas 2.6% and 5.2% of students moved from pass to fail with global scores and total percentage scores, respectively.

Implications of findings

Objective structured clinical examinations and other standardised clinical examination formats forgo the authenticity of observation in clinical practice in order to achieve comparable and fair assessments. This focus on standardisation is important both to reassure the public that a common standard has been achieved and maintain learners' trust in the fairness of examinations. Ensuring that scores adequately reflect performance is paramount to the validity of assessments.[41,42] The utility of assessments emanates from a compromise between several features and no assessment reaches perfect reliability.[43] Nonetheless, the practice of conducting multiple parallel versions of ostensibly the same OSCE examination can be seen to introduce a rarely examined, but important, source of construct irrelevant variance, which has the potential to influence categorisation of a substantial subset of students. Although prior work has attempted to estimate the reliability of fully nested OSCEs,[24] estimation remains difficult and Cronbach's alpha is often used as a surrogate measure.[44] Notably, Cronbach's alpha is not capable of illustrating variance that arises as a result of examiner cohorts, suggesting that other methods are needed to monitor (and potentially adjust for) the influence of different examiner cohorts.

Assumptions regarding the origins of examiner-cohort variability are critical to interpreting these findings, specifically whether such variance represents a random or systematic influence on students' scores. Classical test theory views assessor variance as random,[45] suggesting that examiner-cohort effects might disappear with greater sampling or reduced error variance.[8] Although the reliability of the OSCE was less than ideal (α=0.62), it was similar to the average value of α=0.66 determined by meta-analysis of other OSCEs,[44] suggesting that our findings have ecological validity. Nonetheless, this emphasises the importance of

examiner training, benchmarking and clear marking criteria to ensure acceptable reliability of OSCEs, especially when used for summative assessment. Equally, as reliability is often influenced more by station specificity than by examiner variability, increasing the number of stations is likely to produce larger increases in reliability than examiner-focused approaches.[8]

Conversely, many medical schools run OSCEs across multiple geographically dispersed sites,[18,46] in which the examiners at each site are drawn from clinicians who practise locally and who rarely interact with clinicians from other sites. In this (very common) instance it is reasonable to suggest that examiner cohorts could be systematically different in their practice norms and beliefs, the cohorts of trainees to whom they are exposed, their specialty mixes and their level of specialisation. In such instances, it is more plausible that differences in examiner cohorts might represent a systematic effect, which might persist despite increased numbers of stations or examiners. Such differences would be especially relevant to examinations conducted between multiple institutions or national examinations. The geographical differences suggested by Sebok et al.[23] in Canada and the regional variations in standard setting for knowledge tests observed in the UK[47] both hint at the potential for important systematic variations of this kind, which have the possibility to negatively impact fairness or even patient safety. Consequently, developing a means to measure (and potentially adjust for) differences between examiner cohorts may help to support the validity of multi-circuit, distributed or national OSCE examinations.

Some prior studies[48,49] have attempted similar analyses by using data from several successive iterations of the examinations and relying on natural movement of examiners to producing partial crossing. Such methods implicitly assume that examiner effects are fixed over time (in both studies around 1 year) and ignore examiner by station interactions. Substantial examiner by station interactions have previously been demonstrated,[50] whereas Harik et al.[51] showed that the utility of estimates of examiner differences markedly reduces by 5–6 months after the estimates are made. We believe that VESCA represents an improvement on these methods as it uses both: (i) station-specific comparator data, and (ii) estimates of examiner effects generated within a few hours of the assessment. It may be possible to design assessments in a manner that enables sufficient crossing or linkage through overlapping examiners, thereby foregoing the need for video-based performances. Although such solutions may be feasible within a constrained locality, they are unlikely to be possible within a distributed examination, without making similar assumptions about either station specificity or examiner stability over time.

Limitations

Despite these strengths, the study has some limitations. Use of video-based performances to achieve partial crossing (or linkage) in the data assumed that examiners' scoring of video performances was representative of their live scoring tendencies. The lack of systematic differences between video and live scores is reassuring, but future work should investigate cognitive or social implications of video-based judgements to support this assumption. Analyses viewed the stringency or leniency of examiner cohorts as fixed effects, thereby ignoring examiner × student interactions,[52] or rater idiosyncrasy,[53] which could make the model less dependable, especially at the level of individual students. In particular, the relatively low reliability implies that the degree of residual (random) error was fairly substantial. As the analyses could only adjust for fixed effects, this limits their dependability and includes the possibility that in some instances adjustment may have made the scores less accurate. Although further empirical or simulation work is undoubtedly required to understand this possibility, such concerns need to be viewed in the context of the threat to validity that examiner stringency or leniency is already known to pose.[54] A small minority of students (6.0%) fitted the MFRM poorly. As a result, their adjusted scores should be interpreted with caution. Analysis did not model other factors known to influence examiners' scoring, for example, rater drift[55] or contrast effects.[56,57] Future developments of the model might seek to estimate and incorporate the influence of these effects.

We modelled the influence on students' classification around artificial cut-scores. Notably, the number of students whose classification is altered by the adjustment is dependent on where the cut-score falls within the distribution of students' ability. As a result, a higher or lower cut-score is likely to produce different results. Nonetheless, in line with the developmental intent of the research, these findings illustrate the potential influence of score adjustment on students' classification.

Estimates of the influence of examiner groups relied on the partial crossing provided by the video-based

performance. It is unclear whether the two videos each participating examiner scored (a maximum of 20 videos per cohort) were sufficient to ensure dependable linkage, particularly for examiner cohorts with lower participation rates. It is likely that a larger number of performances is required to produce dependable estimates. Further empirical or simulation work is needed to determine sampling requirements and the extent to which estimates are improved by greater numbers of videos. The common-comparator video performances were selected on convenience; purposive selection of videos showing disparate levels of students' performance might have improved the diagnostic precision of the model. Students who featured in the videos were volunteers, which could limit the generalisation of their performances to the wider student cohort, although (as described in the results) the distribution of students' abilities in the videos appears to have been broadly representative of wider student ability.

Our methods assumed that videoing students' performances in the OSCE had no influence on either students' or examiners' behaviour. Although the unobtrusive ceiling-mounted positions of the cameras may have mitigated any such effect, future investigation should consider whether cameras increase students' test anxiety[58] or alter examiner behaviour.

Lastly, although the procedures we have developed aim to measure and adjust for differences between examiners, they are not capable of accounting for any other systematic differences between parallel circuits of the examination. If, for example, simulated patients in one circuit portrayed cases in a manner that made them more difficult for students than simulated patients in other circuits, the VESCA procedures would neither measure nor adjust for that effect and estimates of students' abilities in that circuit would tend to be inappropriately reduced.

Future study

Future research should seek to address the limitations described. Empirical work is required to understand and optimise the filming process to ensure the best presentation of information to examiners, and to develop filming methods that are adequately unobtrusive but cost-effective enough to use at scale. Study of the sampling requirements (i.e. the number of videos required to create adequate crossing) and the extraneous influences (random error, rater drift and contrast effects) on the dependability of modelling, as well as the relative merits of different modelling approaches, is needed to enhance the technique. Qualitative or theory-driven research should explore users' perceptions of being filmed, as well as the impact of the intervention on assessment behavior, and the acceptability to students, staff, patients and members of the public of adjusted scores for assessment purposes.

CONCLUSIONS

We have developed a novel collection of processes to estimate and adjust for the influence of different examiner cohorts in fully nested, multi-circuit OSCEs. Pilot use of the technique suggests that examiner cohorts can have a substantial influence on the scores of a significant minority of students and could potentially influence categorisation of around 6.0% of students. Although institutions should rely primarily on assessment design (including sufficient sampling and examiner training) rather than post hoc adjustment to ensure adequate reliability in summative OSCEs, development and validation of VESCA may offer a valuable means to compare standards of assessment judgements between geographically dispersed locations or in national examinations.

Contributors: PY conceived the study and substantially contributed to planning, development, data collection, analysis and manuscript drafting. NC substantially contributed to planning, development and data collection, and contributed to analysis and manuscript drafting. AH substantially contributed to data collection, and contributed to manuscript drafting. HB substantially contributed to data collection, and contributed to manuscript drafting. GMcC substantially contributed to analysis of the data and to manuscript drafting. MH contributed to analysis of the data and to manuscript drafting.
Acknowledgements: we would like to thank Kirsty Hartley, Undergraduate Assessments Manager, School of Medicine, Keele University, for her help in organising the practicalities of collecting data within the context of Year-3 OSCE, the information technology team at Keele University School of Medicine for their help with video filming and processing, and all the examiners and students who took part.
Funding: PY is funded by the National Institute for Health Research (NIHR) Clinician Scientist Award. This article presents independent research funded by the NIHR. The views expressed are those of the author(s) and not necessarily those of the National Health Service, the NIHR or the Department of Health.
Conflicts of interest: none declared.

Ethical approval: ethical approval was granted by Keele University Ethical Review Panel, reference ERP175.

REFERENCES

1 Watling CJ. Unfulfilled promise, untapped potential: feedback at the crossroads. *Med Teach* 2014;**36**: 692–7.
2 Wass V, van der Vleuten CPM, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001;**357** (9260):945–9.
3 Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach* 2011;**33** (6):478–85.
4 Ten Cate O. Entrustability of professional activities and competency-based training. *Med Educ* 2005;**39** (12):1176–7.
5 Ginsburg S, van der Vleuten CPM, Eva KW, Lingard L. Cracking the code: residents' interpretations of written assessment comments. *Med Educ* 2017;**51** (4):401–10.
6 Frank JR, Snell LS, Ten Cate O, *et al*. Competency-based medical education: theory to practice. *Med Teach* 2010;**32** (8):638–45.
7 Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ* 2004;**38** (2):199–203.
8 Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med* 1990;**2** (2):58–76.
9 Swanson DB, van der Vleuten CPM. Assessment of clinical skills with standardized patients: state of the art revisited. *Teach Learn Med* 2013;**25** (1 Suppl):S17–25.
10 Gormley GJ, Hodges BD, McNaughton N, Johnston JL. The show must go on? Patients, props and pedagogy in the theatre of the OSCE. *Med Educ* 2016;**50** (12):1237–40.
11 Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39** (3):309–17.
12 Eva KW. On the generality of specificity. *Med Educ* 2003;**37** (7):587–8.
13 Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;**73** (9):993–7.
14 Newble DI, Swansons DB. Psychometric characteristics of the objective structured clinical examination. *Med Educ* 1988;**22** (4):325–34.
15 Van der Vleuten CPM, van Luyk SJ, van Ballegooijen AMJ, Swanson DB. Training and experience of examiners. *Med Educ* 1989;**23** (3):290–6.
16 Pell G, Homer MS, Roberts TE. Assessor training: its effects on criterion-based assessment in a medical context. *Int J Res Method Educ* 2008;**31** (2):143–54.
17 Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J* 1975;**1** (5955): 447–51.
18 Yeates P, Sebok-Syer SS. Hawks, Doves and Rasch decisions: understanding the influence of different cycles of an OSCE on students' scores using Many Facet Rasch Modeling. *Med Teach* 2017;**39** (1):92–9.
19 Tamblyn RM, Klass DJ, Schnabl GK, Kopelow ML. Sources of unreliability and bias in standardized-patient rating. *Teach Learn Med* 1991;**3** (2):74–85.
20 De Champlain AF, MacMillan MK, King AM, Klass DJ, Margolis MJ. Assessing the impacts of intra-site and inter-site checklist recording discrepancies on the reliability of scores obtained in a nationally administered standardized patient examination. *Acad Med* 1999;**74** (10 Suppl):S52–4.
21 Reznick R, Smee S, Rothman A, *et al*. An objective structured clinical examination for the licentiate: report of the pilot project of the Medical Council of Canada. *Acad Med* 1992;**67** (8):487–94.
22 Floreck LM, de Champlain AF. Assessing sources of score variability in a multi-site medical performance assessment: an application of hierarchical linear modeling. *Acad Med* 2001;**76** (10 Suppl):S93–5.
23 Sebok SS, Roy M, Klinger DA, de Champlain AF. Examiners and content and site: Oh My! A national organization's investigation of score variation in large-scale performance assessments. *Adv Health Sci Educ Theory Pract* 2015;**20** (3):581–94.
24 Swanson D, Johnson K, Oliveira D, Haynes K, Boursicot KAM. Estimating the reproducibility of OSCE scores when exams involve multiple circuits. Paper presented to AMEE 2013 Annual Conference *Colouring Outside the Lines*. 24-28 August 2013, Prague, Czech Republic; Prague Congress Centre.
25 Lefroy J, Gay SP, Gibson S, Williams S, McKinley RK. Development and face validation of an instrument to assess and improve clinical consultation skills. *Int J Clin Skills* 2011;**5** (2):115–25.
26 Pell G, Homer M, Fuller R. Investigating disparity between global grades and checklist scores in OSCEs. *Med Teach* 2015;**37** (12):1106–13.
27 Lefroy J, Roberts N, Molyneux A, Bartlett M, Gay S, McKinley R. Utility of an app-based system to improve feedback following workplace-based assessment. *Int J Med Educ* 2017;**31** (8):207–16.
28 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**1** (8476):307–10.
29 Bond T, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 2nd edn. New York and East Sussex: Routledge 2012.
30 Linacre JM. *Facets Computer Program for Many-Facet RASCH Measurement*. Beaverton, Oregon: Winsteps.com 2017. Available from: https://www.winsteps.com/facets.htm [Accessed 18 December 2018.]
31 Linacre JM. What do Infit and Outfit, Mean-square and Standardized mean? [Internet]. Rasch.Org

website. [cited 2018 Jun 12]. Available from: https://www.rasch.org/rmt/rmt162f.htm
32 R Core Team. Computing R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing 2017. Available from: https://www.r-project.org/.
33 Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015;**67** (1):1–48.
34 Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from Generalized Linear Mixed-effects Models. *Methods Ecol Evol* 2013;**4**:133–42.
35 Bartoń K. MuMIn: multi-model inference. R package. Version 1.42.1. 2018. https://cran.r-project.org/web/packages/MuMIn/index.html [Accessed 28 November 2018.]
36 Grömping U. Relative importance for linear regression in R: the package relaimpo. *J Stat Softw* 2006;**17** (1):1–27.
37 Albano AD. Equate: an R package for observed-score linking and equating. *J Stat Softw* 2016;**74** (8):1–36.
38 Carifio J, Perla R. Resolving the 50-year debate around using and misusing Likert scales. *Med Educ* 2008;**42** (12):1150–2.
39 Norman G. Likert scales, levels of measurement and the "laws" of statistics. *Adv Health Sci Educ Theory Pract* 2010;**15** (5):625–32.
40 Glass GV, Peckham PD, Sanders JR. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev Educ Res* 1972;**42** (3):237–88.
41 Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003;**37** (9):830–7.
42 Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas* 2013;**50** (1):1–73.
43 Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Heal Sci Educ* 1996;**1** (1):41–67.
44 Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ* 2011;**45**:1181–9.
45 Streiner D, Norman G. *Health Measurement Scales*, 4th edn. Oxford, UK: Oxford University Press 2008.
46 Fuller R, Homer M, Pell G, Hallam J. Managing extremes of assessor judgment within the OSCE. *Med Teach* 2017;**39** (1):58–66.
47 Taylor CA, Gurnell M, Melville CR, Kluth DC, Johnson N, Wass V. Variation in passing standards for graduation-level knowledge items at UK medical schools. *Med Educ* 2017;**51** (6):612–20.
48 Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ* 2008;**13** (5):617–32.
49 McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ* 2006;**6**:42.
50 Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore DE. A comparison of physician examiners and trained assessors in a high-stakes OSCE setting. *Acad Med* 2005;**80** (10 Suppl):S59–62.
51 Harik P, Clauser BE, Grabovsky I, Nungester RJ, Swanson D, Nandakumar R. An examination of rater drift within a generalizability theory framework. *J Educ Meas* 2009;**46** (1):43–58.
52 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;**36** (10):972–8.
53 Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Adv Heal Sci Educ* 2013;**18** (3):325–41.
54 Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ* 2014;**48** (11):1055–68.
55 McLaughlin K, Ainslie M, Coderre S, Wright B, Violato C. The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Med Educ* 2009;**43** (10):989–92.
56 Yeates P, O'Neill P, Mann K, Eva KW. Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. *JAMA* 2012;**308** (21):2226–32.
57 Yeates P, Moreau M, Eva K. Are Examiners' judgments in OSCE-style assessments influenced by contrast effects? *Acad Med* 2015;**90** (7):975–80.
58 Harrison CJ, Könings KD, Schuwirth L, Wass V, van der Vleuten C. Barriers to the uptake and use of feedback in the context of summative assessment. *Adv Health Sci Educ* 2014;**20** (1):229–45.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Appendix S1. Configuration of parallel lanes and circuits of Day 1 of the OSCE. Figure illustrates: (i) the filmed subset of students; (ii) the eight separate examiner cohorts, and (iii) the timing of video scoring.

Received 22 June 2018; editorial comments to authors 14 August 2018; accepted for publication 7 November 2018