Title:

Measuring the impact of examiner variability in a multiple-circuit Objective Structured Clinical Exam (OSCE)

Authors:

Peter Yeates, MRCP, PhD

Alice Moult, MSc, PhD

Natalie Cope, PhD

G McCray MA MRes PhD

Eleftheria Xilas, MBChB

Tom Lovelock,

Nicholas Vaughan,

Dan Daw

Richard Fuller, MA, FRCP

Robert K McKinley, MD, FRCGP, FRCP

Corresponding Author:

Peter Yeates, Lecturer in Medical Education Research, School of Medicine, David Weatherall Building, Keele University, Keele, Staffordshire, ST5 5BG, UK.

Tel: 01782 734744. Fax: n/a. Email: [p.yeates@keele.ac.uk](mailto:p.yeates@keele.ac.uk), Twitter: @KeeleMERG

Author Sentence style IDs:

Peter Yeates is a senior lecturer in medical education research, in the School of Medicine at Keele University, Keele, Staffordshire UK, and a consultant in acute and respiratory medicine at Fairfield General Hospital, Pennine Acute Hospitals NHS Trust, Bury, Lancashire.

ORCID: https://orcid.org/0000-0001-6316-4051

Alice Moult is a research assistant in medical education in the School of Medicine at Keele University, Keele, Staffordshire UK.

ORCID: https://orcid.org/0000-0002-9424-5660

Natalie Cope is a lecturer in clinical education (Psychometrics) in the School of Medicine at Keele University, Keele, Staffordshire, UK.

Gareth McCray is a researcher at in the School of Primary, Community and Social Care at Keele University, Keele, Staffordshire UK

Eleftheria Xilas, is a foundation year 1 doctor who has recently graduated from the School of Medicine at Keele University, Keele, Staffordshire, UK.

Tom Lovelock is ICT services manager within the Faculty of Medicine & Health Sciences, Keele University, Keele, Staffordshire UK.

Nicholas Vaughan a senior application developer, directorate of digital strategy and IT services at Keele University, Keele, Staffordshire UK.

Dan Daw is an ICT systems development engineer, in the School of Medicine at Keele University, Keele, Staffordshire, UK.

Richard Fuller is the deputy dean, School of Medicine, University of Liverpool, Liverpool, UK.

ORCiD: https://orcid.org/0000-0001-7965-4864

RK (Bob) McKinley is an emeritus professor of education in general practice, in the School of Medicine at Keele University, Keele, Staffordshire, UK.

ORCiD: https://orcid.org/0000-0002-3684-3435

Abstract:

Purpose: Ensuring examiners in different parallel circuits of Objective Structured Clinical Exams (OSCEs) collectively judge to the same standard is critical to the chain of validity. Recent work suggested that the examiner-cohort (i.e. the particular group of examiners) a candidate meets could significantly alter outcomes for some candidates. Despite this, examiner-cohort effects are rarely examined as fully-nested data (i.e. no cross-over between the students judged by different groups of examiners) limit comparisons. This study aimed to replicate and further develop a novel method called Video-based Examiner Score Comparison and Adjustment (VESCA), so that it can be used to enhance quality assurance of distributed or national OSCEs.

Method: Six volunteer students were filmed on all12 stations in a summative OSCE. In addition to examining live performances, examiners from all 8 separate examiner-cohorts collectively scored the same pool of video performances. Examiners scored videos specific to their station. Video scores provided linkage within otherwise fully-nested data, enabling comparisons by Many Facet Rasch Modelling. Analysis primarily compared and adjusted for examiner-cohort effects. Additional analyses compared examiners’ scoring when videos were embedded (interspersed between live candidates within the OSCE) or judged later via the internet.

Results: Having accounted for differences in students’ ability, different examiner-cohorts scores for the same ability of student ranged from 18.57(68.7%) to 20.49(75.9%), Cohen’s d=1.3. Score adjustment changed up to 16% of students’ classification (pass-fail/ fail-pass) depending on modelled cut score. Internet and embedded video scoring showed no difference in mean scores or variability. Examiners’ accuracy did not deteriorate over the 3 week scoring period allowed for internet-based scoring.

Conclusions: Examiner-cohorts produced a replicable significant influence on OSCE scores which was unaccounted for by typical assessment psychometrics. VESCA offers a promising means to enhance validity and fairness in distributed OSCEs or national exams, whilst internet-based scoring may enhance VESCA’s feasibility.

Despite assessment innovations which claim greater authenticity and impact on learning1–3, Objective Structured clinical Exams (OSCE)4 remain a cornerstone of most assessment systems in medical education because of their apparent ability to ensure consistency and fairness5,6, with corresponding assurance to patients and the public in resulting licencing decisions7,8. Consequently, whilst the validity of OSCEs is evidenced by a chain including station design, blue-printing, rating scale characteristics and examiner training, thoroughly understanding examiner variability remains critical to the resulting validity argument9. This paper describes developments to a novel method to enhance this understanding.

Part of the original validity argument for OSCEs relied on the premise that all students would meet all examiners as they rotated around the OSCE circuit, thereby negating examiner differences10. With rising student numbers, many institutions run multiple parallel, simultaneous OSCE circuits, sometimes distributed across different geographical sites, with different examiners in each circuit. As each student only encounters a single “examiner-cohort”11 (i.e. the particular group of examiners allocated to the specific circuit and time they are examined), any collective difference in judgements between examiner-cohorts which impacts on pass-fail decisions could importantly challenge the fairness and validity of the exam.

Psychometric analyses of OSCEs typically ignore the influence of examiner-cohorts because fully nested OSCE designs (no crossover exists between the students seen by different groups of examiners) make them impossible to analyse. Prior work exploring this phenomenon has suggested that differences in examiner-cohorts accounted for 4.4% of the difference between distributed locations in a UK medical school11, whilst standardized patient-raters at different sites accounted for 0.0%-15.7% of score variance between 21 sites in the USA12 and between 2.0%-17.1% of score variance across 6 sites in the Medical College of Canada National Assessment Collaboration Examination for International Medical Graduates13. This suggests potentially significant effects due to assessor scoring variance at an international level.

Whilst these studies hint at the potential for important influences on distributed or even national licencing exams, they are methodologically limited by the examiner nesting already described5,14, in that they require assumptions that student ability is similar on average across sites, based on a model of true randomisation of candidates to sites and circuits. In 2019, Yeates et al15 reported pilot work on a novel methodology called Video-based Examiner Score Comparison and Adjustment (VESCA). In addition to usual examining, this intervention involved all examiner-cohorts scoring a common pool of videos of students’ performances from the OSCE, thereby creating partial crossing and enabling analysis of otherwise fully-nested examiner-cohort effects. Their pilot work demonstrated a large difference (Cohen’s d=1.06) between the average ratings of some cohorts of examiners. If replicated in distributed summative OSCEs or national licencing exams, this effect could importantly influence outcomes for many candidates, thereby challenging the validity of the assessment, with consequent impacts for candidates, patients and institutions as a result of potentially ‘incorrect’ pass fail decisions.

The primary purpose of this study was to replicate the effects reported by Yeates et al15 on a different sample of students and examiners. We addressed two secondary aims: firstly, to extend Yeates et al’s15 work by investigating how score adjustment influenced students’ classification for a range of different cut scores (rather than the single arbitrary cut score they reported); and secondly to develop their methods by comparing the influence of different video-based scoring methods on examiners’ scores and participation rates. In so doing, we aimed to further develop this method with the goal of enhancing quality assurance of distributed OSCEs or national licencing exams.

**Methods:**

We used Video-based Score Comparison and Adjustment (VESCA)15 to compare and adjust for the influence of different examiner-cohorts within a multi-circuit OSCE. Theoretically this intervention adopts a stance of “examiners as fallible”16, seeking to equate for residual examiner differences after faculty development, training and benchmarking have been used to maximal effect to aid standardisation. VESCA utilises 3 sequential processes: 1/ a small sample of candidates are unobtrusively videoed whilst performing on all stations within a real OSCE exam; 2/ examiners mark the live student performances and are invited to score common station-specific comparator videos. Examiners from each cohort collectively score the same videos, creating score linkage between otherwise fully nested groups of examiners; 3/ statistical analyses are used to compare and adjust for examiner influences. To increase the linkage between examiner-cohorts, we asked examiners to score 4 videos of the station they examined, rather than the 2 videos scored previously15. Four videos represented a balance between increased linkage and the time demands of an intervention we hope can be used in routine OSCEs.

**Setting:**

We gathered data from the Year 3 summative OSCE at Keele University School of Medicine undergraduate degree programme, in April-May 2019. Year 3 is the first predominantly clinical year of the 5 year undergraduate programme. The OSCE comprised 12 stations, including consultation, physical examination and procedural skills. Each station integrated a range of communication, diagnostic reasoning, and practical (technical/procedural) skills. Cases were portrayed by scripted simulated or real patients. Students were judged by trained examiners, who were either medical doctors or clinical skills tutors. Examiners training comprised a video-based benchmarking exercise and a pre-OSCE briefing on the scoring format. Performances were scored on Keele’s domain-based rating scale called GeCoS17, providing a score out of 27 points for each station (see appendix 1 for an example). The OSCE was conducted over 3 days, with 4 stations per day. All students attended on all 3 days in order to complete all 12 stations. Four separate parallel circuits of the OSCE were conducted, repeated in the morning and afternoon, with different students and occasionally overlapping examiners’. This gave eight separate examiner-cohorts. See figure 1 for a schematic of the OSCE and research design.

**Population, recruitment and sampling:**

Participation in the study was voluntary for all individuals. All students, examiners and simulated or volunteer real patients in the OSCE were eligible and were invited to take part by email. Filming was conducted in a single circuit during the first rotation of each day, with a maximum capacity of 6 students in total. Filming volunteers were allocated in order of agreement. Written consent was obtained from all participants who retained the right to withdraw up until the end of filming. All examiners were also invited to score videos and provided consent electronically. Ethical approval was granted by Keele Ethical Review Panel(ERP2413).

**Filming OSCE performances:**

Performances on each station were filmed simultaneously by 2 unobtrusively positioned CCTV cameras (one wide angle, one zoomed). The OSCE otherwise proceeded exactly as normal for these students. The lead author (PY) selected camera angles to ensure adequate images of key components of the task. Videos in which key performance elements were obscured were omitted and otherwise the first 4 of the available videos were selected for each station for video based scoring by examiners. Videos were processed to give simultaneous side-by-side views of procedural and physical examination stations, whilst (in order to minimise processing requirements) a single view was selected for consultation (history, clinical reasoning or patient management) stations.

**Video scoring by examiners:**

We compared two separate methods of video-scoring by examiners, termed “internet” and “embedded”. All examiners from the morning session of each day of the OSCE were invited to participate by visiting a password protected web-site and score four videos specific to the station they had already examined (internet). The website provided a downloadable PDF of the station-specific examiner information and scoring rubric. Examiners entered scores for these video performances via a form on the website. Participants had 3 weeks from the OSCE to complete scoring. As a result, internet scoring was highly flexible, and avoided additional time demands during the OSCE.

Examiners in the afternoon session of each day of the OSCE were invited to score the same 4 station-specific videos “embedded” within their OSCE examining circuit. To achieve this, four gaps (similar to the impact of a rest station) were created within the rotation of students in the afternoon, and participating examiners were provided with tablet computers in order to view and score performances. Whilst this ensured that videos were scored in close proximity to live judgements, it required an additional 40 minutes per OSCE session. In both conditions, examiners were asked to avoid extraneous distractions, to give videos their full attention, scoring them as they would in the OSCE. To ensure similarity with live observations, examiners were asked to watch each video once only, and neither pause or rewind. As in Yeates 201915, examiners who judged the filmed performances live also scored them again via video, in order to enable comparison of their video and live scores. We aligned data from live and video performances prior to analyses.

**Analysis**

We addressed our primary aim to compare and adjust the influence of examiner-cohorts using Many Facet Rasch Modelling (MFRM)18. We used ratings on the 27-point rating scale as the dependent variable, and modelled facets of 1/ student, 2/ examiner-cohort and 3/ station. Analyses were conducted using FACETS v3.82.3 (Winsteps, Western Australia). We performed several analyses to ensure the validity of the MFRM analysis. Firstly, we examined for bias between live and videos scores within the subset of performance for which both were available, using a Bland Altmann plot19. This was performed using the ‘BlandAltmanLeh’ package in RStudio: Integrated Development for R. (RStudio, Inc., Boston, MA). Next, we examined how well data met the assumptions of MFRM: we examined the progression of Rasch-Andrich thresholds20 for each rating scale category to determine whether scores were ordinal and the fit of data to the model for each facet,using Linacre’s fit criteria, (mean square infit/outfit 0.5-1.5=good fit21). Next, as MFRM requires unidimensional data, we performed principal components analysis (PCA) on model residuals to exclude any additional dimensions (or factors) in thedata.

Continuing with our primary aim, we used adjusted scores supplied by FACETS to calculate each student’s score adjustments (i.e. that student’s raw average score minus their adjusted average score). From these we calculated the percentage of students’ whose score adjustments were ≥0.5 standard deviations of student ability.

To understand the potential impact of score adjustment on students’ categorisation we examined the proportion of students whose adjusted score placed them in a different pass/fail category to their raw score. As any such analysis is likely to depend on the proportion of students who fail based on their raw score, and the failure rates of OSCE exams vary between exams, we modelled this impact for OSCEs with a range of failure rates. We modelled all possible cut scores at intervals of 0.1 score points, from 16.0 (59.3%), at which 0% of students failed on their raw score, to 19.0 (70.3%), at which 33.4% of students failed on their raw score. We then calculated the percentage of students whose classification (pass-fail or fail-pass) changed for each of these cut scores.

Moving to our secondary aim, we determined the feasibility and influence of the two methods of video scoring (internet vs. embedded). First we examined response rates between both methods. Next we compared 1/. mean scores and 2/. error variances between the two modalities. Error variances compared whether examiners’ scores showed greater variability in one condition than the other. To do this we used a mixed effects regression model which corrected for student ability (random effect) and station difficulty (fixed effect), and modality (internet vs embedded, fixed effect) using the package “lme4” within RStudio: Integrated Development for R. (RStudio, Inc., Boston, MA). We examined beta coefficients for each modality to determine whether scores differed and used an F test to compare error variances between modalities.

Within the internet scoring condition, we investigated whether examiners’ scores varied more from the mean (i.e. became less accurate) as the interval increased between the OSCE and the examiner scoring their videos. To do this, we first calculated the difference between each score given to each video by each examiner compared to the mean of the scores given to that video by all examiners. We expressed this as a mean absolute error (MAE), by calculating:

Where *e* = an examiner’s score for a video – mean of all examiners’ scores for that video.

We then examined the relationship between videos’ MAE and the hours elapsed since the OSCE, using linear regression. This analysis was conducted in IBM SPSS Statistics for Windows v21.0 (IBM corporation, Armonk, NY).

**Results**

**Summary data:**

One hundred and thirteen students completed the OSCE. Eight students volunteered for filming; only the first six were included in filming due to capacity limitations. No filmed volunteers withdrew and all 12 stations were included. Students’ unadjusted average ability across all 12 stations were normally distributed, ranging from 16.0 points (59.3%) to 22.7 points (84%) out of 27 (mean score 19.5 (72%), standard deviation 1.43 (5.3%)). By comparison, videoed students’ unadjusted average abilities ranged from 17.2 (63.6%) to 21.3 (78.9%), indicating their ability was broadly representative of the student cohort. No other details of filmed students were collected. Seventy-three out of ninety-six (76%) examiners chose to participate in video scoring. Participating and non-participating examiners’ scores for live performances were similar (participating examiners’ mean= 19.4, SD=3.8; non-participating examiners’ mean= 19.9, SD=3.2). Video scores comprised 17.7% of all data.

**Ensuring data adequacy for MFRM analysis:**

Comparison of Live and Video scores for the subset of examiners who scored the same students by both methods indicated that there was no systematic difference between scores produced by the two modalities (average difference 0.16, 95% CIs -1.52 to 1.85, see figure 2). Rasch-Andrich threshold progression for the 27 point rating scale showed disordered thresholds for points 8 and 11 on the rating scale (i.e. point 8 received a lower logit value than point 7, rather than the expected increase in logit value). Cumulatively these accounted for less than 2% of observations. All remaining logits values increased progressively suggesting the scale behaved with sufficient ordinality for the intended analysis.

Data fitted the Many Facet Rasch Model well. All examiner-cohorts and stations showed excellent or good fit, whilst 99% of students showed good model fit or acceptable model fit (either overfit, or mild underfit). One student showed potentially degrading underfit (i.e. their data had potential to distort the model in a manner which, at least theoretically, might bias the estimates for other students); omitting this student altered students’ ability estimates by <0.1%, so we continued with the complete data. Fit parameters are displayed in table 1. Principal components analysis of residuals did not indicate any remaining independent factors, providing evidence of unidimensionality. Collectively these findings supported the use of MFRM analysis.

**Influence of examiner-cohorts on students’ scores**

The relative influences of the three modelled facets (candidates, stations, examiner-cohorts) are shown in figure 3. Adjusted “fair-scores” showed that stations varied in difficulty from 16.6 (61.5%) to 21.9 (81.1%). As all students performed on all stations, this had no systematic influence on students’ scores. Having adjusted for the ability of the students which each group of examiners encountered, adjusted “fair-scores” for examiner-cohorts also varied between examiner cohorts, from 18.6 (68.8%) for examiner cohort 2 (the most stringent, “hawkish” examiner cohort) to 20.5 (75.9%) for examiner cohort 1 (the most lenient, “doveish” examiner-cohort). These fair scores can be interpreted as the average scores these groups of examiners would have given to students of the same ability. As the standard deviation of student ability was 1.43, this difference represents a Cohen’s d=1.3, a large effect.

Adjusting students’ scores for the influence of their examiner-cohorts produced substantial changes for many students: the mean score adjustment was 0.58 points (2%), but 45 students (40%) had a score adjustment ≥0.72 score points (Cohen’s d≥0.5), with the largest adjustment of 1.03 score points (3.8%, Cohen’s d=0.72). The number (and percentage) of students who were reclassified ranged from 1 student (1%) at a cut score of 16.0 (59.3%) to 18 students (16%) at a cut score 18.7 (69.3%). The median number of re-classified students was 6 (5%), which occurred at 17.2 (63.7%).

**Comparison of “Internet” vs “Embedded” video scoring methods**

Examiner participation rates varied by scoring modalities: 43(90%) of examiners invited to score “embedded” videos participated compared with 30(63%) of examiners invited to score videos via the internet. There was no statistically significant difference between performance scores when videos were scored “embedded” (Mean=19.02, SD=4.10) versus via the “internet” (Mean=19.15, SD=4.11) with the adjusted difference being 0.50 marks (95%CIs -0.48 to 1.48, t=1.229(270), p = 0.22). Similarly, there was no statistically significant difference in variance of residuals via either modality (embedded 3.03; internet 3.47 F=0.763(167, 119), p = 0.108), suggesting the extent of examiner variability did not differ across scoring modalities. In the internet condition, no significant relationship existed between the Mean Absolute Error (MAE) for individual examiners’ video scores (i.e. the extent of examiners’ inaccuracy, compared with the mean for that video) and the elapsed time since the OSCE (β=0.00 (95% CI’s -0.003 to 0.004), t=0.14, R=0.013, R2<0.001, p=0.88), suggesting that examiners’ scoring did not become more variable over the 3 week period they were allowed to score internet-based videos.

**Discussion**

**Summary of findings**

We have replicated the findings of Yeates et al15, again showing that the difference between examiner-cohorts can be large in some instances (Cohens d=1.3), and could result in change in pass/fail outcomes for between 0%-16% of students. A range of data was supportive of the VESCA method: video-live score comparisons, scale parameters, dimensionality and fit statistics. Internet-based video scores were similar to embedded scores potentially offering a more flexible means to facilitate scoring.

**Implications of findings**

The difference we have observed between the highest and lowest scoring examiner-cohorts is consistent with the upper end of estimates by Sebok et al13 and Floreck et al12 and larger than the influences of many of the previously reported influences on assessors’ judgements16,22,23 or the impact of examiner training24–26. The implications of these differences for the validity of OSCEs depends on the assessments’ purpose9: examiner-cohort effects might barely impact candidates in an OSCE with a low failure rate used only for pass/fail decisions, whereas a significant minority of candidates may be disadvantaged in an OSCE with a 30% failure rate which is also used to rank candidates.

Critical to the interpretation of these findings is whether differences between examiner-cohorts are viewed as a random or systematic influence. If random, then a longer OSCE with greater reliability may ameliorate the observed differences; if systematic then differences would persist within a more reliable exam. Given known variations in national standard setting for knowledge testing items27 and suggested differences between sites in large-scale performance based exams12,13 there is reason to expect that systematic differences may occur between sites in large scale exams which would therefore persist in otherwise highly reliable OSCEs.

We suggest that routine measurement of examiner-cohort effects is needed in large scale OSCEs. Whilst methods based on differential rater functioning28 give valuable insights into some examiner biases, they are unlikely to be informative for fully-nested OSCE designs. VESCA promises a feasible means of achieving this without assuming examiners are stable entities over long time intervals29 or across different stations30. Given the extent of known examiner variability31, debate is needed in our community over the merits of score adjustment based on psychometric analyses.

Both embedded and internet methods of video-scoring by examiners produced similar scores for videos. Embedded scoring produced considerable resource demands: videos had to be processed and made available quickly and numerous assistants were required to supply examiners with the correct tablets at the right times. Additionally it made the afternoon session longer by 40 minutes. Given these constraints, internet scoring may be more realistic within usual practice.

**Limitations**

Despite the rigour of our study it has some limitations. Modelling relied on the linkage provided by four videos per participating examiner (double that in Yeates 201915) and resulted in 17.7% linkage in the data. Whilst more than double the 8% minimum linkage required in work by Linacre32, caution is required as Linacre’s method optimally balanced the design whilst our method extrapolated from partial crossing. As a result greater linkage might potentially have produced different estimates.

Many Facet Rasch Modelling only models consistent differences between examiners; if examiners’ scoring was influenced by rater drift33,34 contrast effects35,36, examiner x student interactions37 or idiosyncracy38 then these effects would not be adjusted. Fit of data to the model was generally good, but dependability of score adjustment would be limited for less well fitting students. Examiners were not randomised to embedded vs internet scoring. Consequently we cannot exclude the possibility that a difference due to modality was obscured by an unknown confounding effect.

**Future research**

Further research, both empirical and using simulation, is needed to determine: the accuracy of adjusted scores; how examiners’ stability influences modelling; and the likely impact of operational variables such as the number of linking videos and the choice of statistical analysis method. Research should investigate whether the demographic characteristics of students in the videos (including age, sex, ethnicity) influence examiners’ scoring. VESCA constitutes a complex intervention; qualitative exploration should explore stakeholders’ reactions, behaviours, preferences and trust in relation to VESCA to enable it to be implemented within existing assessment culture.

**Conclusions**

Score differences between examiner-cohorts appear to be a significant and replicable effect. Routine consideration should be given to these effects in distributed OSCEs as part of quality assurance procedures. VESCA offers a promising method to do so, which could enhance the validity and trust in distributed or national exams.

*Acknowledgments,*

Many thanks to Kirsty Hartley, Sharon Simpson and the assessments team at Keele School of Medicine, all of the students, examiners and simulated and volunteer patients who took part in the study, and to Keele School of Medicine for allowing and facilitating the study.

*Funding/Support,*

Peter Yeates is funded by a Clinician Scientist Award from the National Institute for Health Research (NIHR). This paper presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

*Data*

All data were obtained from Keele School of Medicine, and have been used with their permission. Whilst the study findings have been presented within Keele School of Medicine the manuscript has not been subject to any requirement for internal review prior to publication. Data has not been placed in a data repository.

*Other disclosures,*

None

*Ethical approval,*

Ethical approval was granted by Keele Ethical Review Panel, reference ERP2413.

*Disclaimers*,

None

*Previous presentations*.

Portions of this work have previously been presented at the Association for Medical Education in Europe (AMEE) annual meeting, Vienna, 2019 and the Ottawa Conference on Assessment of health professionals and evaluation of programmes, Kuala Lumpur, 2020.

Word count: 3551 words

References:

1. Frank JR, Snell LS, Cate O Ten, et al. Competency-based medical education: theory to practice. Med Teach. 2010;32(8):638-645.

2. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ. 2005;39(3):309-317.

3. ten Cate O. Entrustability of professional activities and competency-based training. Med Educ. 2005;39(12):1176-1177.

4. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. 2004;38:199-203.

5. Swanson DB, van der Vleuten CPM. Assessment of clinical skills with standardized patients: state of the art revisited. Teach Learn Med. 2013;25(S1):S17-S25.

6. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Med Teach. 2010;32(10):802-811.

7. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA. 2002;287(2):226-235.

8. Wass V, Vleuten C Van Der, Shatzer J, Jones R. Medical education quartet. Assessment of clinical competence. Lancet. 2001;357:945-949.

9. Kane MT. Validating the interpretations and uses of test scores. J Educ Meas. 2013;50(1):1-73.

10. Harden RM, Stevenson M, Downie WW, Wilson GM. Medical education assessment of clinical competence using objective structured examination. Br Med J. 1975;1:447-451.

11. 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):92-99.

12. 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):S93-S95.

13. 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):581-594.

14. Swanson D, Johnson K, Oliveira D, Haynes K, Boursicot KAM. Estimating the reproducibility of osce scores when exams involve multiple circuits. In: AMEE Annual Conference. Colouring Outside the Lines. Prague, Czech Rebulic; 2013:2F/4.

15. Yeates P, Cope N, Hawarden A, Bradshaw H, McCray G, Homer M. Developing a video-based method to compare and adjust examiner effects in fully nested OSCEs. Med Educ. 2019;53(3):250-263.

16. 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):1055-1068.

17. 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 Ski. 2011;5(2):115–125.

18. Bond T, Fox C. Applying the rasch model fundamental measurement in the human sciences. 2nd Ed. New York & London: Routledge; 2012.

19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-310.

20. Linacre JM. A User’s Guide to FACETS Rasch-model computer programs manual v3.83.0. Rasch.Org website. https://www.winsteps.com/winman/copyright.htm Accessed Sept 19, 2019.

21. Linacre JM. What do infit and outfit, mean-square and standardized mean? Rasch.Org website. https://www.rasch.org/rmt/rmt162f.htm. Accessed June 12, 2018.

22. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. Med Educ. 2016;50(5):511-522.

23. Lee V, Brain K, Martin J. From opening the ‘black box’ to looking behind the curtain: cognition and context in assessor-based judgements. Adv Heal Sci Educ. 2019;24(1):85-102.

24. Eric S. Holmboe, MD; Richard E. Hawkins, MD; and Stephen J. Huot. Effects of training in direct observation of medical residents ’ clinical competence. Ann Intern Med. 2004;140:874-881.

25. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med. 2009;24(1):74-79.

26. 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):143-154.

27. 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):612-620.

28. Wind SA, Sebok-Syer SS. Examining differential rater functioning using a between-subgroup outfit approach. J Educ Meas. 2019;56(2):217-250.

29. 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).

30. Brennan PA, Croke DT, Reed M, et al. Does changing examiner stations during UK postgraduate surgery objective structured clinical examinations influence examination reliability and candidates’ scores? J Surg Educ. 2016;73(4):616-623.

31. Eva KW. Cognitive influences on complex performance assessment: lessons from the interplay between medicine and psychology. J Appl Res Mem Cogn. 2018;7(2):177-188.

32. Linacre JM. Many-facet rasch measurement. 2nd Ed. Chicago: MESA Press; 1994.

33. 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):989-992.

34. Coetzee K, Monteiro S. DRIFT happens, sometimes: Examining time based rater variance in a high-stakes OSCE. Med Teach. 2019;41(7) 819-823.

35. 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):2226-2232.

36. Yeates P, Moreau M, Eva K. Are examiners’ judgments in osce-style assessments influenced by contrast effects? Acad Med. 2015;90(7):975-980.

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

38. 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):325-341.

**Tables and figures**

Table 1: Fit of data within each Facet to the Many Facet Rasch Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Facet | Number (%) | Range of Infit/Outfit Mean-Square values | Range of Infit/Outfit Z-standardised  values | Data fit |
| Examiner-Cohorts | All | 0.7-1.3 | ≤±2.0 | Excellent |
| Stations | All | 0.5-1.5 | -3.44 to 3.65 | Good / Productive |
| Students | 93 (82%) | 0.5-1.5 | - 1.53 to 2.24 | Good / Productive |
|  | 12 (13%) | 0.24-0.5 | -1.45 to -2.68 | Overfit |
|  | 7 (8%) | 1.5-2.0 | <2.0 | Mild underfit |
|  | 1 (1%) | 2.06 | 2.1 | Underfit |

Figure 1: Schematic Overview of Study Design and Research Data Collection

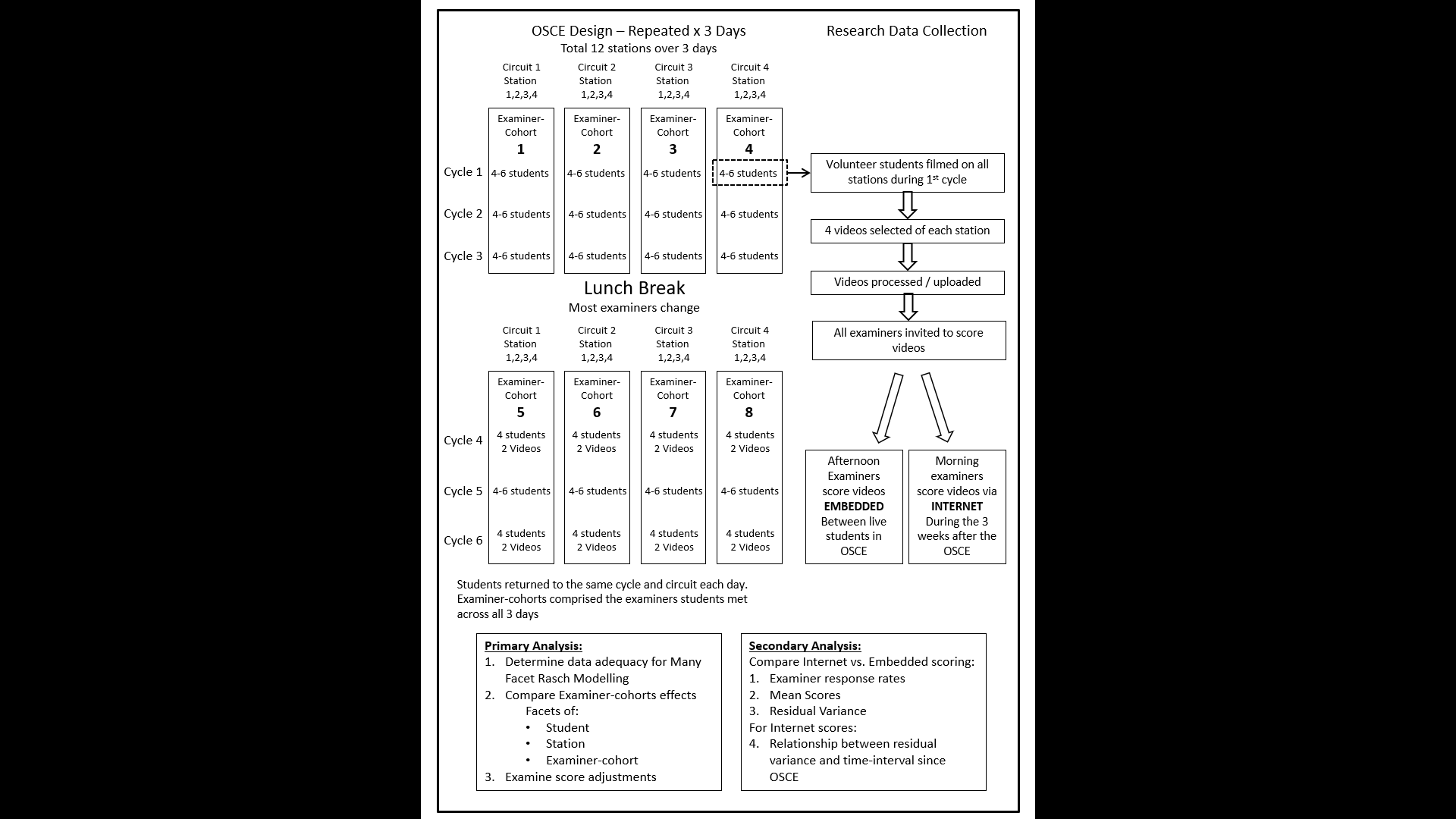
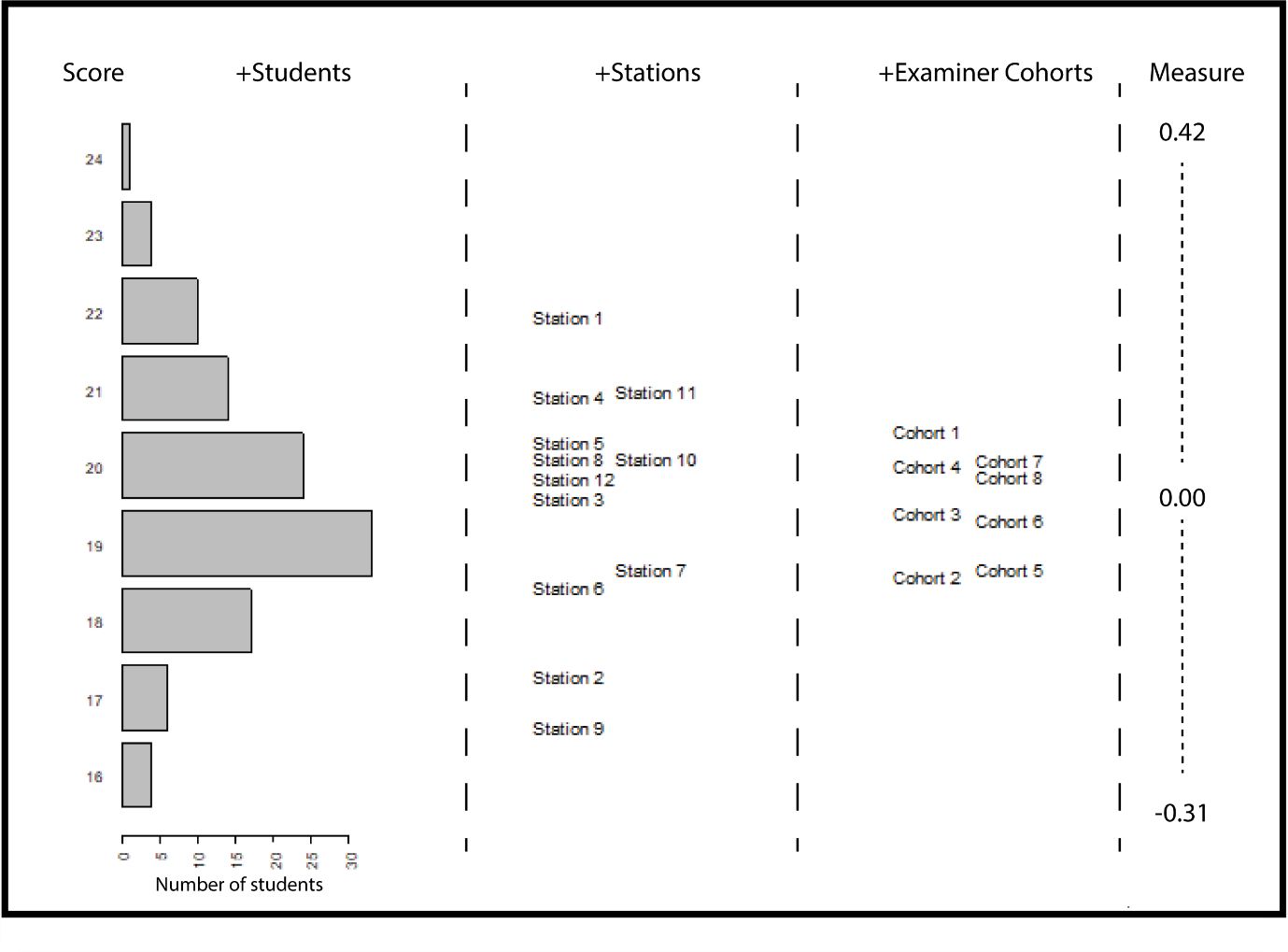


Figure 2: Bland Altman plot of the difference between Live and Video scores (out of 27, vertical axis) plotted against the average of the scores given to video and live versions of each performance (out of 27, horizontal axis) for the subset of examiners who judged the same performances by both methods. Dashed horizontal lines indicate the mean difference between the two measures and 1.96 standard deviations from the mean difference.



Figure 3: Wright Map showing the relative ability/difficulty/stringency of the facets of student, station and examiner cohorts respectively. Horizontal bars in the students’ column indicate the number of students at each level of ability. Examiner-cohorts indicate the score that different examiner cohorts would give to a student of the same ability.



Appendix 1:

**Example of MARKING AND FEEDBACK CATEGORIES from 1 station**

**N.B. Domains varied by station. The 5 most appropriate domains for each station were selected by station authors.**

**This example is from a consultation-focused station. Examiners were required to give a mark 1-4 for each of the listed domains.**

**Please mark the skill categories from 1 – 4:**

1 – Must improve in this category

2 – Borderline in this category

3 – Proficient in this category

4 – Very good in this category

**NB: station specific examples of the content of each domain were provided for examiners in addition to the generic domain descriptors. These have been excluded due to restrictions on sharing of sensitive examination content by the host institution**

**Generic domain descriptors:**

|  |
| --- |
| **History Process** |
| * Enables patient to fully elaborate presenting problem(s) * Skilled use of questioning including open and closed questions * Clarifies words used and/or symptoms presented by patient as appropriate * Recognises and responds appropriately to verbal and non-verbal cues * Listens attentively |
| **History Content** |
| * Obtains sequence of events * Obtains details of symptoms * Obtains effect on the patient's life * Obtains patient's ideas, concerns and expectations * Obtains relevant background information including: Past Medical, Drug, Family and Social History; Systems review; Factors influencing health |
| **Clinical Reasoning** |
| * Seeks relevant and specific information from patient's record or third parties * Generates appropriate working diagnoses or problem list * Seeks discriminating information from history, examination and investigations to help confirm or refute working diagnoses * Correctly interprets information obtained * Applies basic, behavioural and clinical sciences to solution of patient's problem * Recognises limits of competence and acts accordingly |
| **Management Content** |
| * Investigates appropriately * Refers appropriately |
| **Building and Maintaining the Relationship** |
| * Develops and maintains a professional relationship with patient * Respects the patient's ideas, beliefs and autonomy * Responds empathically * Fosters collaboration |

**Examiners were additionally required to provide an overall global score for the performance.**

**Global Score Descriptors:**

|  |  |
| --- | --- |
| **DESCRIPTOR** |  |
| Demonstrates inadequacies in all categories, and one or more serious defects.  *Practice likely to result in harm to the patient or self or other staff* | **1** |
| Demonstrates inadequacies in many categories, and one or more serious defects.  *Practice could result in harm to the patient or self or other staff* | **2** |
| Demonstrates inadequacies in several categories. No serious defects.  *Practice unlikely to result in harm to the patient, self or other staff* | **3** |
| On balance, capable in several relevant categories to a satisfactory standard. Any deficiencies are minor  *Competent and Safe practice* | **4** |
| On balance, capable in many relevant categories to a satisfactory standard. Any deficiencies are minor  *Competent and Safe practice* | **5** |
| Capable in all relevant categories to a high standard.  *Skills and attitude are above the level expected for student in this module* | **6** |
| Capable in all categories to a very high standard.  *Skills and attitude are well above the level expected for student in this module* | **7** |

**Based on these domains, the tablet-based mark system required that examiners gave each candidate a score in each domain:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Domain 1: History Process | 1 | 2 | 3 | 4 |  |  |  |
| Domain 2: History Content | 1 | 2 | 3 | 4 |  |  |  |
| Domain 3: Clinical Reasoning | 1 | 2 | 3 | 4 |  |  |  |
| Domain 4: Management Content | 1 | 2 | 3 | 4 |  |  |  |
| Domain 5: Building and Maintaining the Relationship | 1 | 2 | 3 | 4 |  |  |  |
| Global Judgement: | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

**The score on each domain plus the global judgment rating were** **summed to give a score out of 27.**