VESCA’s variable precision: Determining the accuracy of adjustment for examiner differences in distributed OSCEs

Peter Yeates (✉ p.yeates@keele.ac.uk)  
Keele University  https://orcid.org/0000-0001-6316-4051

Gareth McCray  
Keele University  https://orcid.org/0000-0002-0728-5171

Research Article

Keywords: Assessment, Equivalence, Psychometrics, Simulation, Many Facet Rasch Modelling

Posted Date: May 4th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2889959/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Introduction: Ensuring examiner equivalence across assessment locations is a priority within distributed Objective Structured Clinical Exams (OSCEs) but is challenging due to lack of overlap in performances judged by different groups of examiners. Yeates et al have develop a methodology (Video-based Examiner Score Comparison and Adjustment (VESCA)) to compare and (potentially) adjust for the influence of different groups of examiners within OSCEs. Whilst initial research has been promising, the accuracy of the adjusted scores produced by VESCA is unknown. As this is critical to VESCA's utility, we aimed to investigate the accuracy of adjusted scores produced by VESCA under a range of plausible operational parameters.

Methods: using statistical simulation, we investigated how: 1/proportion of participating examiners, 2/number of linking videos, 3/baseline differences in examiner stringency between schools, 4/number of OSCE stations and 5/different degrees of random error within examiners’ judgements influenced accuracy of adjusted scores. We generated distributions of students’ “true” performances across several stations, added examiner error, and simulated linking through crossed video-scoring, before using Many Facet Rasch Modelling to produce adjusted scores, replicating 1000 times for each permutation, to determine average error reduction and the proportion of students whose scores became more accurate.

Results: Under all conditions where no baseline difference existed between groups of examiners (i.e. random rather than systematic variance), score adjustment minimally improved or worsened score accuracy. Conversely, as modelled (systematic) baseline differences between schools increased, adjustment accuracy increased, reducing error by up to 71% and making scores more accurate for up to 93% of students in the 20% baseline-difference condition.

Conclusions: score adjustment through VESCA will substantially enhance equivalence for candidates in distributed OSCEs when 10–20% baseline differences exist between examiners in different schools. As such differences are plausible in practice, consideration should be given to use of VESCA in large scale/national exams.

Introduction

Contemporary views of assessment have evolved over recent years to emphasise longitudinal low stakes assessment to support learning(1). Despite this, Objective Structured Clinical Exams (OSCEs)(2) remain prominent for ensuring learners’ competence prior to progression between phases or for licensure decisions because of their ability to test under controlled conditions, according to a defined blueprint and their apparent ability to ensure equivalence of testing standards(3).

Validity of an assessment relates to its intended purpose(4). Thus, whilst fairness in assessments can be conceptualised in different ways(5), within summative OSCEs, ensuring that candidates face equivalent testing is critical to the validity argument(6). Examiner variability is a well-established phenomenon across all forms of performance assessment(7) and so resultantly poses potential challenges to
equivalence. Prior research has shown that examiner variability within OSCEs can be substantial\(^8\)–\(^{10}\) thereby producing construct irrelevant variance which can threaten the equivalence and thereby the validity of summative interpretation of OSCE scores.

Within large scale summative OSCEs a further problem arises. Owing to student numbers, OSCEs are typically distributed across multiple parallel tracks of (ostensibly) the same exam, or indeed across widely spaced geographical locations. This poses a further challenge, as there is the potential that examiners in different locations may hold systematically different frames-of-reference when judging performance\(^{11}\). Whilst this potential has been illustrated by a limited range of prior work\(^{12},^{13}\) it is rarely studied as assessment designs are typically fully nested (i.e. no crossover between the students seen by different groups of examiners) which makes it challenging to explore these effects. Recently Yeates et al\(^{14}\) have developed a method called video-based examiner score comparison and adjustment (VESCA) which specifically aims to overcome this challenge. VESCA employs three sequential phases: 1/ a sample of candidates are videoed on each of the stations within the OSCE; 2/ all examiners, in addition to judging live candidates, are asked to score a number of videos of student performances from the station they examined; 3/ the partial crossing created by the video-scores is used to link different examiner groups (termed “examiner-cohorts”) within statistical analyses to compare and equate for examiner effects.

Yeates et al have used VESCA within a number of studies, in each case showing differences between the estimates of different examiner-cohorts ranging from 5.7\%\(^{14}\), 6.9\%\(^{15}\) to 7.1\% \(^{16}\). Importantly, resulting score adjustments suggested that a proportion of students would vary their pass/fail classification (up to 16\% depending on cut score \(^{16}\)) or their rank position \(^{15}\). Critically, as the authors acknowledge in each paper, these observations depend on a strong assumption that the adjusted scores produced by VESCA (generally through Many Facet Rasch Measurement) are indeed more accurate representations of candidates’ true performance than their raw scores. Yeates et al\(^{17}\) used subset re-sampling from Yeates 2021 data to explore this potential. By varying the number of linking videos per participating examiner and the proportion of examiners who scored videos, they showed that candidates’ score adjustments (i.e. the difference between their adjusted and raw scores) were sensitive to changes in both of these parameters. The purpose of this current study is to extend that work, by determining the accuracy of the adjusted scores produced by VESCA, and to explore how that accuracy varies under a plausible range of different operational parameters.

Several parameters could conceivably influence the accuracy of score estimates produced by VESCA. As Yeates et al\(^{17}\) have previously shown that firstly examiner participation rates and secondly the number of linking videos scored by each examiner can both influence score adjustments, both these variables seem germane to understanding VESCA’s accuracy. Theoretically, we would expect that with greater amounts of linking material (i.e. more videos per station; greater examiner participation), the Many Facet Rasch model would develop more accurate estimates of examiner-cohort effects as the impact of random variability on these estimates would be reduced. Third, OSCEs frequently vary in their number of constituent stations, which is a significant influence on reliability\(^{18}\). Consequently, station numbers
could influence VESCA's accuracy. Fourth, Many Facet Rasch Modelling can adjust for systematic variations between examiners, but prior research has shown that a significant proportion of score variance in OSCEs is random or unexplained (19). As a result, determining the impact of different levels of random variability on VESCA's score adjustments is important.

Lastly, the express purpose of VESCA is to compare examiners' influence across distributed sites where examiners and students are “nested” together (i.e. no crossover between the candidates seen by different groups of examiners). Prior research has suggested that inter-site variations may account for up to 16–17% of score variance in some instances (12, 13). Further work has suggested that examiners' frame of reference relates to the typical standard of performance to which they are exposed (20). Consequently, it is conceivable that location A could have highly capable candidates and stringent examiners, whereas location B could have less capable candidates and lenient examiners. Notably, whilst examiner equivalence would be highly different between these locations, unadjusted OSCE scores could be very similar. We refer to this potential systematic difference in examiner stringency between different institutions effect as “examiner baseline differences”. As its ability to adjust for these effects is critical to the intended use of VESCA, we additionally sought to understand the influence of baseline differences on the accuracy of VESCA score adjustments.

To operationalise “accuracy” in practical terms we considered 1/ the proportion of candidates whose scores became more accurate and 2/ the reduction in total error variance in each scenario. We then asked the following research questions:

1. How is the accuracy of score estimates produced by VESCA influenced by:
   a. The number of linking videos per examiner (0, 2, 4, 6, or 8 linking videos)
   b. The proportion of examiners who participate in scoring videos (50%, 65%, 80%, 100%)
   c. The combination of these 2 effects.

2. How is the accuracy of score estimates produced by VESCA influenced by:
   a. Differing extents of baseline differences in examiner stringency between different sites (0%, 5%, 10%, 20%)
   b. The number of stations in the OSCE (6, 12, or 18 stations)
   c. The combination of these two effects

3. How is the accuracy of score estimates produced by VESCA influenced by reduction in the degree of random variability in examiners' scoring (random error divided by 2, by 4, and by 8)

We considered one further issue as an ancillary research question. Given that some degree of imprecision is inevitable in all statistical modelling, it seemed plausible that the accuracy of adjusted scores may
relate to the size of the adjustment being made, i.e. large adjustments may be more accurate than small adjustments, because the ratio of size of the adjustment to the size of the imprecision may be greater (i.e. a greater signal to noise ratio). This may enable a score adjustment threshold to be determined above which score adjustments reach greater accuracy. We examined this within all data produced by studies 1–2, by asking:

4. How does the proportion of candidates whose scores become more accurate vary for different sizes of score adjustment for each of the parameters investigated within RQs 1–2.

Methodology

Adopting a post-positivist stance, which holds that a singular mind-independent truth exists which is only partially knowable through enquiry (21), we opted to address these questions through statistical Monte-Carlo simulation (22). We chose this because it enables researchers to specify an underlying truth, then to sequentially simulate disruption of that truth, and procedures which aim to recover the truth, and then the ability to precisely compare the recovered and originally specified truths, in order to determine how closely they match. While we accept that a statistical simulation may oversimplify empirical data in some situations, we posit that this is the best approach to addressing these questions as in all other situations, the distance from the underlying truth (and therefore accuracy) is unknown (23).

Simulated data generation

We simulated the operation of VESCA through three sequential processes (see Fig. 1), in which we simulated and combined the effect of several known influences on OSCE scores. All parameter estimates were derived from empirical data, based on re-analysis of data from Yeates 2021 (16).

Firstly, we modelled the “true” performance of a range of students on each station in an OSCE using a simple sum-score approach. Data were generated using the GeCos scale (24) which was used in Yeates at al (16). This combines ratings on several performance domains to give a scale minimum of 6 and maximum of 27. To do this we randomly generated a distribution of students’ overall ability (M = 19.47 out of 27; SD = 1.13 (5.4% of scale)) and then generated a range of station difficulties (SD 1.52 (7.2% of scale)) and an idiosyncratic student x station interaction (SD = 1.71 (8.1% of scale)). We combined these, using a linear function to produce students’ simulated “true” performance on each station in the OSCE.

Secondly, we added examiner variability to these scores by creating a distribution of examiners (SD = 1.40 (6.7% of scale)). Examiners were randomly allocated to a station and to 1 of 4 examiner cohorts (i.e. distinct groups of examiners) such that each students’ “true” scores were exposed to a unique group of examiners stringencies, and the same examiner stringency applied to all students for a given station within a cohort. As examiners did not change station, we could not model examiner x station effects. Next, we simulated an additional random error term (SD = 2.35 (11% of scale)) to capture additional unmodeled variation in examiners’ scoring (for example due to the time of day (25), contrast (26) or halo (27) effects
from the previous candidates, examiner-student interactions, and any other unknown sources of variability). We summed the students’ “true” performance score on each station, with the examiner stringency and the additional random term to give the student’s “observed score” on each station in the OSCE – the score they would have actually received on each station in the OSCE.

Formally, generation of the observed students’ scores can be expressed as:

$$\text{Score}_{ijk} = \beta_0 + u_1 \text{Station}_i + u_2 \text{Student}_k + u_3 \text{Student: Station}_{ik} + u_4 \text{Examiner}_j + \epsilon_{ijk}$$

Where: $\beta_0$ the overall model intercept (i.e, average student score in the dataset), $u_1$ station difficulty $i$, $u_2$ student ability $k$, $u_3$ the interaction between student $k$ and station $i$, $u_4$ examiner $j$ stringency, and $\epsilon_{ijk}$ is the residual error.

Thirdly, we mimicked the influence of the VESCA procedure by randomly selecting a specified number of student performances on each station and nominating these as “video performances”. A proportion of examiners were then randomly selected to “participate” (see RQ 3) and the stringency values of these examiners + the random error term were applied to the relevant “video performances” for the station they had examined. This created an additional set of crossed “video scores” for each station as would be collected by using VESCA (i.e. the same “video performances” were scored by multiple examiners from different examiner cohorts). This created a dataset comprised of students’ “live” observed scores on each station in the OSCE, and further observed video scores allocated to station-specific videos by examiners. All data generation was performed in R.

As in the procedures used by Yeates et al(16), these data were then analysed using Many Facet Rasch Modelling, in FACETS(28) to produce an adjusted overall (i.e. average) score for each student (see Fig. 1). This is because FACETS only outputs an overall adjusted score for each student, rather than a score for each station for each student. Transferred data were labelled with student ID, examiner cohort, and station and included both live and video scores. These data were transferred to FACETS using the R package “immer”(29). We ran each simulation 1000 times in order to obtain stable estimates. As this was computationally demanding, simulations were run via 16 virtual machines on a 16-core server each linking R to facets using the R package “immer”(29). A schematic overview of the data simulation process can be seen in Fig. 1.

Computationally, generated data via a flexible function written in R(30). The function always has four cohorts of examinees but allows the manipulation of i) the number of linking videos, ii) the min and max of the score range, iii) the numbers of stations, iv) the number of candidates, v) the number of cohorts, vi) the number of examiners, vii) the mean ability of a candidate, viii) the standard deviation of candidate scores, ix) the standard deviation of station difficulties, x) the standard deviation of examiner stringencies, xi) the standard deviation of a station by candidate interaction (i.e., the error in the ‘performance score’) and xii) the expected proportion of examiners who would participate in the linking process.

Simulations
Several simulations were conducted to mimic the effects of the VESCA method in various contexts. Unless otherwise specified, simulations modelled 12 stations, 60 students in 4 cohorts with 48 examiners, with an assumed 80% of examiners participating, and 4 linking videos.

**Study 1** – The first study addressed RQ1 by modifying the number of linking videos (0, 2, 4, 6 and 8) and the expected proportion of examiners to consent to providing linking data (50%, 65%, 80% and 100%). This included modelling “typical” conditions (i.e. Yeates et al 2021) which comprised 4 linking videos and 80% participating examiners. No baseline differences between schools were modelled in study 1. All permutations of parameter values were simulated for a total of 5 (range of linking videos) x 4 (range of examiner participation rates) = 20 sets of 1000 simulations for each unique pair of values.

**Study 2** – The second study addressed RQ2, by looking at the effect of changing the number of stations (6, 12, 18) and the degree of site-related baseline difference in examiner stringency / student leniency (0%, 5%, 10%, 20%) – see last paragraph of background for definition. Baseline differences were modelled selecting 2 examiners-cohorts as “school A” and 2 examiner cohorts as “school B” and then adding or subtracting the relevant percentage score to the students and examiners coefficients for each school. We assumed that examiner stringency was completely negatively correlated with student ability (i.e., as students became more able, examiners were more stringent and thus the mean expected scores between sites would be equal). All possible combinations of parameter values were simulated for a total of 3 (numbers of stations) x 4 (degrees of baseline difference) = 12 sets of 1000 simulations for each unique pair of values.

**Study 3** – The third study examined RQ3 by reducing the size of the overall residual error term on the performance of the VESCA linking model. This was done by dividing the error term by 2 (error/2 – i.e. 50% of error in prior studies); by 4 (error/4, 25% of the error in prior studies) or by 8 (error/8, 12.5% of the error in prior studies). The objective of this study was not to investigate a plausible real-life situation (as reducing the residual error is a very difficult thing to achieve in practice) but to understand the impact that this residual score error was having on the functioning of VESCA.

**Measurement of performance**

Having generated data under each of these parameters and used them to obtain FACETS estimates of each students’ adjusted score, we used them to determine accuracy of the estimates.

To do this, we calculated three variables for each student, for all 1000 iterations of each permutation of each study:

A. Observed Score Error: The mean absolute difference (MAD) of the observed score – the performance score. This gave the residual error of each student's observed score, from their “true” score, prior to adjustment.

B. Adjusted Score Error: The mean absolute difference (MAD) of the adjusted score – the performance score. This gave the residual error of each student's score, from their “true” score, after adjustment via the
For the VESCA method to show utility, we would expect the adjusted scores to be closer to the “true” scores than the observed scores. Lastly, we calculated:

**C. Score Adjustment**: The mean absolute difference of the adjusted score – the observed score

This gave the size of the adjustment made to each student’s score using the VESCA method.

We then calculated the first of our dependent variables: the proportion of students whose adjusted score became more accurate than their observed score (for brevity, termed “pAcc”). This was defined as the proportion of students for whom “adjusted score error” < “observed score error” (i.e. VESCA score adjustment had resulted in a score nearer to their “true” performance score).

For each permutation of each study, we then calculated:

1. The mean of all students “observed score error”
2. The mean of all students “adjusted score error”
3. The ratio of mean “adjusted score error” : mean “observed score error” (i.e. 1 / 2.)

This demonstrated, on average, how much score accuracy changed for each permutation in each study. For brevity, we term this the “error ratio” (ErR), noting that values below 1 indicated improved accuracy and values above 1 indicated reduced accuracy.

To address RQ4 (how does the proportion of candidates whose scores become more accurate vary for different sizes of score adjustment), we categorised each students’ data in each permutation of each study, based on the size of the score adjustment they received, using categories of score adjustment (expressed as a percentage of the assessment scale) of: [0–1%), [1–2%), [2–3%), [3–4%), [4–5%), [5–6%), [6–7%), [7–8%), [8–9%), > 9%. Next, we further categorised students based on the extent of change in the accuracy of their adjusted scores compared to their observed scores (i.e. how much more or less accurate their adjusted score became), using categories also based on percent of the assessment scale of (<-6%), (-6%—-4%], (-4%—-2%], (-2%—0%], [0%-2%], [2%—4%), [4%-6%], (> 6%). We then tabulated these results for inspection. To aid categorisation of these findings, we used a target of 80% of students’ scores becoming more accurate in order to define whether a useful threshold could be established.

**Results**

Data were generated in 35 separate simulations, resulting in 25,200,000 “performance” scores (i.e. scores for 2,100,000 students on an average of 12 stations).

Research Question 1: How is the accuracy of score estimates produced by VESCA influenced by:
a. The number of linking videos per examiner (0, 2, 4, 6, or 8 linking videos)

b. The proportion of examiners who participate in scoring videos (50%, 65%, 80%, 100%)

c. The combination of these 2 effects

These questions were addressed by study 1. The accuracy of adjusted scores across all parameters modelled in this study were low. Notably, this study assumed that there were no baseline differences between examiners in different sites. Error ratio (ErR) values ranged from a worst case 1.22 (i.e. adjusted scores contained 22% more error than observed scores) for 2 linking videos, with 50% examiner participation, to a best case of 0.94 (i.e. score adjustment removed 6% of the error in the observed scores) for 8 linking videos with 100% examiner participation. The proportion of students whose scores became more accurate (pAcc) as a result of adjustment corresponded closely, ranging from pAcc = 0.44 (44% of students’ scores became more accurate; 56% of students’ scores became less accurate) for 2 linking videos / 50% examiner participation, to pAcc = 0.53 (53% of students’ scores became more accurate) for 8 linking videos / 100% examiner participation. A detailed breakdown of all permutations of these parameters can be seen in Table 1.
Table 1
Influence of number of linking videos per examiner and proportion of participating examiners on adjusted score accuracy

<table>
<thead>
<tr>
<th>Number of Linking Videos per Examiner</th>
<th>Proportion of participating examiners</th>
<th>Mean Error in Observed scores (SD)</th>
<th>Mean Error in Adjusted Scores (SD)</th>
<th>Error ratio</th>
<th>Proportion of students' whose scores became more accurate through adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>50</td>
<td>0.603 (0.46)</td>
<td>0.623 (0.47)</td>
<td>1.03</td>
<td>0.48</td>
</tr>
<tr>
<td>0</td>
<td>65</td>
<td>0.605 (0.46)</td>
<td>0.619 (0.47)</td>
<td>1.02</td>
<td>0.48</td>
</tr>
<tr>
<td>0</td>
<td>80</td>
<td>0.605 (0.46)</td>
<td>0.618 (0.47)</td>
<td>1.02</td>
<td>0.49</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>0.600 (0.45)</td>
<td>0.618 (0.47)</td>
<td>1.03</td>
<td>0.48</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>0.597 (0.45)</td>
<td>0.728 (0.56)</td>
<td>1.22</td>
<td>0.42</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
<td>0.587 (0.45)</td>
<td>0.676 (0.52)</td>
<td>1.15</td>
<td>0.44</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>0.588 (0.45)</td>
<td>0.643 (0.5)</td>
<td>1.09</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>0.589 (0.45)</td>
<td>0.612 (0.47)</td>
<td>1.04</td>
<td>0.48</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>0.584 (0.45)</td>
<td>0.661 (0.52)</td>
<td>1.13</td>
<td>0.44</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
<td>0.580 (0.45)</td>
<td>0.618 (0.48)</td>
<td>1.07</td>
<td>0.47</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>0.579 (0.45)</td>
<td>0.592 (0.46)</td>
<td>1.02</td>
<td>0.49</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>0.579 (0.45)</td>
<td>0.565 (0.44)</td>
<td>0.98</td>
<td>0.52</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>0.573 (0.44)</td>
<td>0.625 (0.49)</td>
<td>1.09</td>
<td>0.46</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>0.569 (0.44)</td>
<td>0.586 (0.46)</td>
<td>1.03</td>
<td>0.48</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>0.570 (0.45)</td>
<td>0.563 (0.44)</td>
<td>0.99</td>
<td>0.50</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>0.563 (0.45)</td>
<td>0.538 (0.43)</td>
<td>0.96</td>
<td>0.52</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>0.567 (0.44)</td>
<td>0.614 (0.48)</td>
<td>1.08</td>
<td>0.46</td>
</tr>
<tr>
<td>8</td>
<td>65</td>
<td>0.563 (0.45)</td>
<td>0.569 (0.45)</td>
<td>1.01</td>
<td>0.49</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>0.557 (0.44)</td>
<td>0.544 (0.43)</td>
<td>0.98</td>
<td>0.51</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>0.556 (0.45)</td>
<td>0.524 (0.42)</td>
<td>0.94</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Accuracy of the adjusted scores was independently influenced by both the number of linking videos and the proportion of participating examiners. Changing the number of linking videos per examiner (whilst averaging across all of the included categories of examiner participation, i.e. keeping this constant) gave error ratios for 0 video = 1.03, 2 videos = 1.13, 4 = 1.05, 6 = 1.02, 8 = 1.00, with corresponding proportions of students seeing increased score accuracy (pAcc) values of 0 videos = 0.48, 2 videos = 0.45, 4 = 0.48, 6 = 0.49, 8 = 0.50 respectively. Notably, therefore, the accuracy of adjusted scores was reduced (compared
to no linking) by having 2 linking videos per examiner, but then progressively slowly increased for larger number of linking videos.

Changing the proportion of participating examiners (whilst averaging across all of the included categories of linking videos, thereby keeping those constant) showed a more linear pattern, giving error ratios for 50% of examiners = 1.11, 65% of examiners = 1.06, 80% = 1.02 and 100 of examiners = 0.99. Corresponding proportions of students whose scores became more accurate (pAcc) were 50% of examiners = 0.45, 65% examiners = 0.47, 80% = 0.49 and 100% = 0.51 respectively.

Research Question 2: How is the accuracy of score estimates produced by VESCA influenced by:

a. Differing extents of baseline differences in examiner stringency between different sites (0%, 5%, 10%, 20%)

b. The number of stations in the OSCE (6, 12, or 18 stations)

c. The combination of these two effects

These questions were addressed by study 2. The accuracy of adjusted scores varied substantially in this study. Error ratio (ErR) values ranged from a worst case 1.42 (i.e. adjusted scores contained 42% more error than observed scores) for 0% baseline difference in examiner stringency, with 18 OSCE stations, to a best case of 0.29 (i.e. score adjustment removed 71% of the error in the observed scores) for 20% difference in baseline examiner stringency with 12 OSCE stations. The proportion of students whose scores became more accurate (pAcc) as a result of adjustment showed a corresponding pattern, ranging from pAcc = 0.37 (only 37% of students’ scores became more accurate for 0% baseline difference and 18 OSCE stations, to pAcc = 0.93 (93% of students’ scores became more accurate) for 20% baseline difference and 18 OSCE stations, with a very similar finding (pAcc = 0.92) for 20% baseline difference and 12 OSCE stations. A detailed breakdown of all permutations of these parameters can be seen in Table 2.
### Table 2
Influence of stations in the OSCE and degree of baseline difference in examiner stringency on adjusted score accuracy

<table>
<thead>
<tr>
<th>Degree of baseline difference between school (% of scale)</th>
<th>Number of Stations in OSCE</th>
<th>Mean Error in Observed scores (SD)</th>
<th>Mean Error in Adjusted Scores (SD)</th>
<th>Error ratio</th>
<th>Proportion of students whose scores became more accurate through adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>0.814 (0.63)</td>
<td>0.829 (0.64)</td>
<td>1.02</td>
<td>0.49</td>
</tr>
<tr>
<td>0</td>
<td>12</td>
<td>0.579 (0.45)</td>
<td>0.592 (0.46)</td>
<td>1.02</td>
<td>0.49</td>
</tr>
<tr>
<td>0</td>
<td>18</td>
<td>0.475 (0.37)</td>
<td>0.674 (0.52)</td>
<td>1.42</td>
<td>0.37</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>0.907 (0.7)</td>
<td>0.828 (0.64)</td>
<td>0.91</td>
<td>0.54</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>0.712 (0.54)</td>
<td>0.592 (0.46)</td>
<td>0.83</td>
<td>0.59</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>0.635 (0.47)</td>
<td>0.673 (0.52)</td>
<td>1.06</td>
<td>0.49</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>1.172 (0.85)</td>
<td>0.825 (0.64)</td>
<td>0.70</td>
<td>0.67</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>1.056 (0.68)</td>
<td>0.589 (0.46)</td>
<td>0.56</td>
<td>0.75</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>1.026 (0.59)</td>
<td>0.67 (0.52)</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>2.012 (1.07)</td>
<td>0.82 (0.64)</td>
<td>0.41</td>
<td>0.85</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>1.996 (0.83)</td>
<td>0.586 (0.45)</td>
<td>0.29</td>
<td>0.92</td>
</tr>
<tr>
<td>20</td>
<td>18</td>
<td>1.998 (0.72)</td>
<td>0.667 (0.52)</td>
<td>0.33</td>
<td>0.93</td>
</tr>
</tbody>
</table>

### Table 3
Influence of reduction in examiner random error on adjusted score accuracy

<table>
<thead>
<tr>
<th>Reduction in error</th>
<th>Mean Error in Observed scores (SD)</th>
<th>Mean Error in Adjusted Scores (SD)</th>
<th>Error ratio</th>
<th>Proportion of students whose scores became more accurate through adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error / 2</td>
<td>0.375 (0.29)</td>
<td>0.323 (0.25)</td>
<td>0.86</td>
<td>0.56</td>
</tr>
<tr>
<td>Error / 4</td>
<td>0.302 (0.24)</td>
<td>0.207 (0.16)</td>
<td>0.69</td>
<td>0.62</td>
</tr>
<tr>
<td>Error / 8</td>
<td>0.28 (0.22)</td>
<td>0.166 (0.13)</td>
<td>0.59</td>
<td>0.66</td>
</tr>
</tbody>
</table>
Accuracy of the adjusted scores showed different relationships with the baseline difference in examiner stringency and the number of OSCE stations. Changing the baseline difference in examiner stringency (whilst averaging across the 3 different numbers of OSCE stations, i.e. keeping this parameter constant) gave error ratios for 0% baseline difference = 1.15, 5% baseline difference = 0.93, 10% = 0.64, and 20% = 0.34 with corresponding proportions of students seeing increased scores accuracy (pAcc) values of at 0% baseline difference = 0.45, 5% = 0.54, 10% = 0.71 and 20% = 0.90 respectively. Consequently at 0% baseline difference in examiner stringency, score adjustment made scores less accurate, whereas at 20% baseline difference in examiner stringency, 66% of error was removed and 90% of students’ scores became more accurate.

Changing the number of stations in the OSCE (whilst averaging across all levels of baseline difference in examiners stringency, thereby keeping those constant) gave error ratios for 6 OSCE stations of 0.76, 12 stations of 0.68 and 18 stations of 0.87. Corresponding proportions of students whose scores became more accurate (pAcc) were 6 stations = 0.64, 12 stations = 0.69, and 18 stations = 0.62. Consequently, these different numbers of OSCE stations produced a U-shaped influence on adjusted score accuracy, with adjustments made from an OSCE with 12 stations showing greater accuracy than the score adjustments made from either a 6 or 18 station OSCE. Notably, however, the extent of error in observed scores for 18 stations (i.e. the amount of error contained in the unadjusted scores produced by examiners) is lower than for 12 stations (3rd column Table 2), so this observation may arise from an interaction of the effectiveness of score adjustment with the amount of error originally present.

Research Question 3: How is the accuracy of score estimates produced by VESCA influenced by reduction in the degree of random variability in examiners’ scoring (random error divided by 2, by 4, and by 8)

This question was addressed by study 3. As in study 1, it was performed with an assumption of 0% baseline difference between sites, and used standard parameters (12 station, 4 linking videos and 80% examiner participation). Accuracy of adjusted scores increased progressively as the amount of random error was reduced. Error ratios (ErR) for the usual extent of random examiner error = 1.02, half usual random examiner error (err/2) = 0.86, one quarter random error (err/4) = 0.69, and one eighth usual random examiner error (err/8) was 0.59. Corresponding proportions of students whose scores became more accurate were: usual examiner error = 0.49, err/2 = 0.56, err/4 = 0.62, err/8 = 0.66. Consequently, whilst reducing the degree of random error we modelled within examiners scoring increase accuracy, a very substantial increase in examiners random error (one eighth its usual value) produced a moderate increase in accuracy (41% reduction in error; 66% of students’ scores became more accurate).

Research Question 4: How does the proportion of candidates whose scores become more accurate vary for different sizes of score adjustment for each of the parameters investigated within RQs 1–2

This study produced 32 tables of tabulated results. These findings, along with summary text and further details of how they were calculated, are presented in appendix 1. In summary, when there was no baseline difference between sites (i.e. study 1) the findings did not demonstrate a threshold for any of the studied parameters beyond which the target of pAcc > 0.8 was achieved. Notably the vast majority of
adjustments made in study 1 were comparatively small. When larger baseline differences existed (10–20% baseline difference, see study 2) adjustments were typically larger, with a majority exceeding 9% of the assessment scale for 20% baseline differences. Thresholds in the region of 3–4% of the assessment scale could be set for scenarios where a baseline difference of 20% existed, to achieve a target of pAcc > 0.8. Notably, therefore adjustment thresholds depended on the degree of baseline difference rather than an absolute value of the adjustment threshold.

**Discussion**

**Summary of Results:**

This study has produced several novel insights into the accuracy of score adjustments produced by VESCA under a range of plausible OSCE conditions. Firstly, it has shown that in the typical conditions under which it has previously been used, i.e. 12 stations, with between 2 (14) and 4 (15, 16) linking videos and 50% (15) to 80% (16) examiner participation rates (and assuming that there were no systematic differences between the groups of examiners in the parallel forms of the OSCE), score adjustment overall makes scores less accurate than the observed scores produced by examiners, with an increase in the total error and only a minority of students’ scores becoming more accurate. This finding is surprising and unexpected. Next, it has shown that the accuracy of adjusted scores are indeed sensitive to the theorised parameters of the number of linking video per examiner, and the proportion of examiners who participate, but, within what we expect to be reasonable limits, increasing these parameters only modestly increased the accuracy of the resulting score adjustments. Next, the study has shown that when there are large baseline differences (i.e. 20%) between the stringency of examiners at different sites, VESCA-based score adjustment becomes very much more accurate, reducing error by 59–73% and resulting in 85–93% of students’ scores becoming more accurate, under typical operating conditions (12 stations, 4 linking videos per examiners, 80% examiner participation). Lastly, we have shown that accuracy of adjusted scores can be increased by reducing examiners’ random error variability, but that a substantial reduction in this error (i.e. 1/8th its usual extent) is required to produce a moderate (41%) reduction in error, making 66% of students’ scores more accurate.

**Theoretical considerations:**

Whilst Many Facet Rasch Modelling can be used for a number of purposes, one of its explicit intended applications is to place disparate groups of examiners on a common scale (or within a single frame of reference) by linking and equating for their differences(31). Consequently, whilst VESCA’s processes within OSCEs are comparatively novel, the process of using limited linkage to equate for examiner differences using Many Facet Rash Modelling is not. Indeed, established guidance on using Many Facet Rasch Modelling provides consideration of different linking patterns, including more sparse linkage patterns than we employed in this study (32). None of this material suggests that there are circumstances in which adjusted scores produced by FACETS will become less accurate than the original scores, so some readers may be surprised to see that this occurred to some extent in all of our scenarios and was
very frequent in all situations where there were no baseline differences between sites. This may be because the extent of random variance in OSCEs is comparatively large\(^{(19)}\). From a theoretical perspective, it appears that Many Facet Rasch Modelling becomes increasingly robust when there are progressively larger systematic differences to account for, and when there is less random error. Conversely, stronger linking through greater proportions of examiner participation or more linking videos produces modest improvements in accuracy. Essentially, Many Facet Rasch Modelling (and by extension VESCA) is useful for adjusting when systematic differences are substantially larger than random variability but performs poorly when systematic differences are small relative to random error. Critically, therefore, practitioners who seek to use it to adjust scores need to know when there are large systematic differences between groups of examiners.

Practical Recommendation

Establishing the extent of systematic difference between sites or examiner-cohorts is in most conventional distributed OSCEs is extremely difficult, as observed scores confound the combined influence of student ability and examiner stringency, meaning that observed scores may mask differences in examiner stringency, or conversely that observed differences may arise due to genuine differences in students’ performance. By asking examiners to score station specific videos, VESCA provides controlled comparisons of examiners’ scoring on a subset of the examined scores and are therefore directly applicable to the task in hand. Consequently, we recommend that in practice, the first 2 steps of VESCA (1/ filming videos and 2/ asking examiners to score them) should be implemented in distributed OSCEs, and the scores allocated to videos by different groups of examiners should be directly compared as part of quality assurance procedures. These comparisons should be made within a short period after the OSCE, before results are released. Where these comparisons suggest negligible differences, faculty should be reassured. Where they suggest 5–10% differences, the evidence we have provided here suggests that rather than using adjusted scores, faculty should then focus faculty development efforts on sites or groups of examiners where scoring was discordant, but they should continue to use the original score. If video scores suggest a baseline difference between sites in the region of 10–20% of the assessment scale, then faculty may consider using the adjusted scores instead of raw scores as these will substantially reduce error and will increase score accuracy for the overwhelming majority of students.

Having made this recommendation, it is pertinent to consider how likely it is that such baseline differences will occur in practice. The simplest answer to this question may be that as they are rarely measured, we do not know. Sebok et al\(^{(12)}\), however, attributed up to 17% of observed score variance to examiners in different sites in a national distributed examine, although this occurred for only a minority of occasions examined. Whilst comparing standard setting for knowledge testing, Taylor 2017 et al\(^{(33)}\) found up to 25% points difference between different schools who set the highest and lowest standards for the same items. Consequently, it is clear that large site-based variations can occur in assessment in some instances, and therefore it seems both prudent to monitor for their occurrence and reassuring to know that adjustment can be dependable when differences are substantial.
Limitations

Our study has some limitations. All simulation is limited by the parameters of the simulation. In this study, we modelled all known substantial influences on OSCE scores (candidate, station, examiner, and appropriate random variance terms)(18, 19, 34), but omitted influences shown more recently to be minor such as contrast effects or differential rater function over time (35). Importantly, we can't comment on combinations of parameters which we didn't test (for example 60% examiner participation, 3 linking videos or 12% baseline difference) nor can we infer beyond the range of modelled parameters (i.e. 12 linking videos). None the less, we assert that the modelled parameters represent a realistic range of likely use.

Each simulation only modelled 60 students. This value was chosen for computational simplicity. Adding additional students would have required a greater number of examiner groups, but would not have increased their distribution beyond the specified range in each simulation, so is unlikely to have produced different results.

Our modelling made unidimensional assumptions. Multi-dimensionality in the data could theoretically have further attenuated accuracy. OSCE data examined within prior uses of VESCA have been unidimensional (14–16) and as data dimensionality should be checked before MFRM is used, we assert that this assumption was reasonable.

Future research:

As with all modelling, these data would benefit from independent replication by a different group adopting a different approach. Further investigation could investigate the accuracy of “examiner-cohort” measures (as described in (16)) to determine whether they offer a more accurate measure for quality control purposes. Further research could determine VESCA’s accuracy in some of the scenarios we didn’t test, for example a 15% baseline difference with either 100% or only 65% examiner participation, in order to extend our understanding of when it is reasonable to use adjusted scores in practice.

Conclusion

The accuracy of score adjustment produced by VESCA under typical operating conditions, when there are no baseline differences between examiner groups, is low and we do not support the use of adjusted scores from VESCA under these circumstances. Conversely, when large baseline differences exist between locations, score adjustment becomes substantially more accurate and consideration could be given to using VESCA-adjusted scores in these scenarios. By comparing examiners’ scoring of videos, VESCA provides directly relevant controlled comparisons of the influence of different examiner groups from different locations within distributed OSCE. These findings offer a basis to support its use in practice within defined parameters.

Declarations
Acknowledgements:

none

Ethics:

As this study generated data from statistical simulation, it did not involve human participants or their data. No ethical approval was sought.

Funding:

Peter Yeates is funded through a National Institute for Health Research (NIHR) Clinician Scientist. The study constitutes independent research and does not represent the views of the NIHR, the NHS or the department of health and social care.

Authors’ contribution:

PY developed the initial concept, significantly contributed to the design and method and significantly contributed to the interpretation of the finding, and drafting of the manuscript

GM contributed to the initial concept, and significantly contributed to the design. GM performed all code writing and analysis, and contributed to the interpretation of the finding and to drafting of the manuscript

Both authors approved the final draft

Conflicts of Interest:

None

Data Availability:

Computer code used for generation and analysis available on request.

References


15. Peter Yeates, Adriano Maluf, Natalie Cope, Gareth McCray, Stuart McBain, Dominic Beardow, et al. Using Video-based Examiner Score Comparison and Adjustment (VESCA) to compare the influence of examiners at different sites in a distributed Objective Structured Clinical Exam (OSCE). Do Hawks Nest with Doves? In Submission with Advances in Health Sciences Education [Internet]. 2022 [cited 2022 Oct 14]; Available from: https://www.researchsquare.com/article/rs-1862169/v1


Figures

![Conceptual Diagram](image)

**Figure 1**

A conceptual diagram of the data simulation process.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.