Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: www.tandfonline.com/journals/imte20

Determining the influence of video-based benchmarking (VBB) on examiner variability in objective structured clinical exams (OSCE): The Align study

Peter Yeates, Rebecca Jane Edwards, Aditya Narain, Robert McKinley, Janet Lefroy, Gareth McCray, Giles Roberts, Ellie Hammond, Stu McBain, Andrew Blythe, Kathy Cullen, Craig Napier, Laura Sims, Harish Thampy, Tushar Vince, Sue Ensaff, Rhian Goodfellow, Christopher Harrison, Ching-Wa Chung, Steven Capey, Chris Roberts & Rebecca Vallender

To cite this article: Peter Yeates, Rebecca Jane Edwards, Aditya Narain, Robert McKinley, Janet Lefroy, Gareth McCray, Giles Roberts, Ellie Hammond, Stu McBain, Andrew Blythe, Kathy Cullen, Craig Napier, Laura Sims, Harish Thampy, Tushar Vince, Sue Ensaff, Rhian Goodfellow, Christopher Harrison, Ching-Wa Chung, Steven Capey, Chris Roberts & Rebecca Vallender (01 Mar 2026): Determining the influence of video-based benchmarking (VBB) on examiner variability in objective structured clinical exams (OSCE): The Align study, Medical Teacher, DOI: 10.1080/0142159X.2026.2631743

To link to this article: https://doi.org/10.1080/0142159X.2026.2631743

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2026.2631743

Determining the influence of video-based benchmarking (VBB) on examiner variability in objective structured clinical exams (OSCE): The Align study

Peter Yeates[a], Rebecca Jane Edwards[b], Aditya Narain[a], Robert McKinley[a], Janet Lefroy[a], Gareth McCray[a], Giles Roberts[a], Ellie Hammond[a], Stu McBain[a], Andrew Blythe[c], Kathy Cullen[d], Craig Napier[e], Laura Sims[f], Harish Thampy[g], Tushar Vince[h], Sue Ensaff[i], Rhian Goodfellow[i], Christopher Harrison[j], Ching-Wa Chung[k], Steven Capey[l], Chris Roberts[m] and Rebecca Vallender[i]

[a]School of Medicine, Keele University, Keele, UK; [b]Stoneygate Centre for Empathic Healthcare, University of Leicester, Leicester, UK; [c]Bristol Medical School, University of Bristol, Bristol, UK; [d]School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, UK; [e]School of Medicine, Dentistry and Nursing, University of Glasgow, Glasgow, UK; [f]University of Exeter Medical School, Exeter, UK; [g]School of Medical Sciences, University of Manchester, Manchester, UK; [h]Kent and Medway Medical School, Canterbury, UK; [i]School of Medicine, Cardiff University, Cardiff, UK; [j]School of Medicine, University of Lancashire, Burnley, UK; [k]School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, UK; [l]Medical School, Swansea University, Swansea, UK; [m]School of Medicine and Population Health, The University of Sheffield, Sheffield, UK

ABSTRACT

Introduction: Reducing examiner variability in Objective Structured Clinical Exams (OSCEs) is a priority within clinical performance assessment. In contrast to typical OSCE examiner training, video-based benchmarking (VBB) involves examiners scoring videos a/from their specific station b/shortly before the OSCE and then reflecting on and discussing scores/justifications agreed by an expert panel. Whilst realist evaluation has described mechanisms and contexts by which VBB may operate, VBB's overall efficacy is unknown.

Methods: We performed a multi-centre (12 UK medical schools) stratified randomised controlled trial of VBB versus control to determine the influence of VBB on examiners' score variability and other score characteristics. Secondarily, we compared the average scores allocated by examiners from different schools.

Results: 171 medically qualified, trained OSCE examiners participated in the study. VBB showed no significant effect on overall examiner variability. In pre-specified analyses, VBB reduced variability from group mean of initially 'outlying' examiners on the borderline performance (VBB mean variability 3.02 out of 27 (IQR1.98-4.98), control 4.70 (3.91–5.70), $p < 0.016$) and made examiners more likely to correctly fail a minimally failing performance ($p < 0.03$, OR = 2.133[95% CI 1.081–4.208]). VBB caused a small increase in confidence. There were no significant differences in average scores by school.

Conclusions: VBB may enhance trust in OSCEs through more accurate classification of borderline performances and aligning outlying examiners scoring.

ARTICLE HISTORY
Received 5 November 2025
Accepted 9 February 2026

KEYWORDS
Assessment; OSCEs; examiner variability; randomized control trial; video-based benchmarking

Introduction

Ensuring that new graduates attain the intended standard of clinical performance remains critical to patient safety and the wider missions of health professionals' education [1]. Accurately assessing students' performance is essential to this. Objective Structured Clinical Examinations (OSCEs) [2] remain a backbone of many assessment programmes globally whether within a competence based medical education (CBME) paradigm or a more traditional approach [3]. Generally, where OSCEs are used within assessment programmes, their epistemological assumptions [4] are positivist (i.e. they aim to estimate the 'correct' score for each student) and so it is epistemologically consistent within such situations to minimize variability due to examiners' judgements. By reducing construct-irrelevant examiner variability in OSCEs, the resulting scores will have greater precision [5] and would be expected to more reliably detect weaker performances. As a result, reducing examiner variability may indirectly contribute to enhancing patient safety [6]. Within Kane's model of validity this could be argued to enhance the 'scoring'

CONTACT Peter Yeates p.yeates@keele.ac.uk School of Medicine, David Weatherall building, Keele University, Keele, Staffordshire ST5 5BG.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/0142159X.2026.2631743.

Practice points

- Video-based benchmarking (VBB) aims to align examiners judgements in OSCEs with a pre-specified standard, developed before the OSCE by an expert panel.
- VBB involves asking examiners to score and reflect on and discuss station-specific video performances shortly before an OSCE.
- Whilst VBB did not show an overall reduction in examiner variability, it helped to align outlying examiners on borderline performances and more than doubled the chances of examiners failing a minimally failing performance.
- VBB may offer a means to enhance trust in OSCEs. Sharing benchmarked stations between institutions may help to manage pragmatic demands of implementation.

domain [7], by increasing the extent to which scores reflect the true level of underlying performance, thereby aiding the overall validity of the assessment.

Although considered a cornerstone of good practice within OSCEs [8–10], efforts to reduce assessor variability through training have historically shown only limited benefits. For example, Holmboe et al. found that rater training made assessors more stringent in their scoring but had little impact on the range of their scores [11] whilst Cook et al. found no significant effect of examiner training [12]. Pell and colleagues also found a stringency effect of training in observational data, although didn't investigate examiner consistency [8]. More recently, Kogan et al. produced a moderate reduction in assessor variability through a longitudinal training programme which involved a total of 6 h of training over a 3-month period [13]. Whilst this is encouraging, this degree of training may not always be practical with the number of examiners which are required for large scale OSCEs. Typical OSCE examiner training programmes use generic examples of performance within frame of reference training (FORT) and are generally conducted weeks or even months before the OSCE.

In contrast to these generally disappointing findings on assessor training, assessors' judgements are well known to be influenced by recently viewed performances, through either 'contrast' [14,15] or 'assimilation' [16,17] effects. Whilst these effects have been used to demonstrate that assessors' scoring can be biased, they also illustrate that it is possible to shift assessors frame of reference through a short-term intervention. This correspondingly suggests that by using a different configuration of performances it may be possible to reverse these effects, using them instead to increase the alignment of examiners scoring around a more similar frame of reference. Notably, these studies have typically used a short *interval* between performances and used performances which are *specific* to the task which is due to be assessed. Consequently, this combination (short interval, specific example) may offer a means to align examiners judgements in a way that prior training has not managed.

Given these observations, interest has arisen in procedures which are typically termed 'Video-Based Benchmarking', which is arguably a subset of more traditional frame of reference training [18], but which provides examiners with one or two video-based examples of performance by candidates on the task they are about to examine shortly before examining commences. Anecdotally, a number of institutions within the UK employ similar procedures, but no data is available about their prevalence and there is little research evidence to establish their effectiveness. Recently, Edwards et al. [19] used Realist evaluation to explore how VBB worked differently for different examiners within a range of contexts to produce benefits to perceived alignment. They suggested that well engaged examiners who are provided with credible video-benchmarks are able to locate their judgements relative to an agreed standard and adjust their frame of reference through reflection and discussion. Their study was limited, though, in that whilst examiners perceived their scores would be more aligned, the study was not designed to determine whether examiner variability was reduced as a result. A follow up pilot experimental trial of VBB showed no overall effect of VBB on examiner variability although it was small so may have been under-powered [20]. Post hoc analysis within this study suggested that examiners who were outliers (i.e. their initial scoring was a long way from the group mean), and who received the VBB intervention, showed reduced variability in their scores for the borderline performance compared with similar examiners who had not experienced VBB. Prior research on examiners' judgements has suggested that additional assessment data may increase examiners' confidence without necessarily increasing accuracy [21]. Conversely, examiners who lack confidence may be less willing to allocate failing grades [22]. Consequently, it is of further interest to understand how VBB may influence examiners' confidence in their ability to rate performances.

Given the important potential benefits of video-based benchmarking to OSCE practice, the theoretical plausibility of the intervention and the existence of early empirical descriptions of the contexts and mechanisms by which it may operate, but uncertainty about whether these observations translate

into meaningful reductions in examiner variability, we considered it imperative to study the influence of video-based benchmarking on OSCE examiner variability. To do this, we asked the following research questions:

Does Video-based Benchmarking (VBB):

1. Reduce overall examiner variability
2. Reduce variability of initially outlying examiners
3. Increase the proportion of correctly classified failing performances
4. Increase accuracy (i.e. distance from panel-agreed score)
5. Increase judgmental confidence

As a secondary, incidental focus, given that there has been recent interest in potential inter-medical school variations in examiner scoring within the UK [23], and given that VBB might have the potential to reduce any such variations, we additionally asked:

6. Do examiners from different medical schools differ by school in the average scores they allocate to the same performances, and if so, by how much?
7. If so, does VBB reduce any such inter-school variations in average examiner scores?

Methods

Study design and theoretical assumptions

To address these questions, we performed an online, randomised, multi-centre, stratified, parallel groups, controlled interventional trial [24] of the influence of video-based benchmarking on examiners scoring. Data was collected *via* online workshops which additionally facilitated discussion between participants as part of the intervention. We chose this design because we consider the underlying phenomenon of interest (examiner variability) to have mind-independent existence (it is measurably observable) but cannot be reliably reported by individual examiners (examiners perceptions of the accuracy of their scoring may be very different to their measured accuracy). For this reason, and to complement prior realist work that explored *how* VBB might work in different contexts, we adopted an objectivist method to test *whether* it works at scale, noting this inevitably constrained our focus, in this case to a single OSCE station.

Population, sampling, and recruitment

The target population for this study was medical school OSCE examiners from the United Kingdom. We aimed to sample these representatively by recruiting a geographically diverse sample. Sample size was estimated through an *a priori* power calculation which used an F-test to determine sample requirements to detect changes in variance. No internal correlations were modelled because the study was expected to have a balanced design. This estimated that 64 participants per group were required to have 80% power to detect a 30% 2-tailed reduction in examiner variability with $\alpha = 0.05$. Given uncertainty regarding the sample distribution we therefore set a conservative target of 80 participants per group.

Eleven partner medical schools were recruited *via* collaboration with the Medical Schools Council Assessment Alliance (MSCAA) – 12 schools participated, including the research team's origin school. To be eligible, participating schools had to be GMC approved and have previously run graduation level OSCE exams. Participating schools came from a wide geographic distribution (North, South, and Midlands of England, Scotland and Wales) and included established Russell group and newer medical schools. Partner schools were asked to distribute recruitment advertisements to their OSCE examiner pool *via* a standardised email, aiming for up to 20 examiners per school. The study inclusion and exclusion criteria were:

Inclusion/exclusion criteria

Examiners must be:

- Medically qualified to practise medicine in the UK.
- Trained as an OSCE examiner or previously examined OSCEs in a UK Medical School.

Examiners must not:

- Appear in any of the videos used in this study.

Prospective participants were randomly allocated to intervention or control groups using a random number generator embedded in the survey platform. Randomisation was conducted independently within each school. Blocks of five participants per condition were used as an upper limit to maintain approximate balance of participants in the control and intervention groups within each school.

Ethics

Participation in the study was voluntary. Participants provided consent *via* an online survey and had the right to withdraw. Ethical approval for the study was granted by [name of institution] Research Ethics Committee [0535]. Demographic information was collected on a voluntary basis along with the consent information.

Materials

OSCE station and videos. The study used video-based performances which came from a prior study [23] and featured graduation-level performances in a formative OSCE exam. All videos were from the same station which involved students discussing management of their condition with a patient who has asthma. The station tested information gathering skills, management reasoning, knowledge of the management of asthma, communication skills and shared decision making, and was developed using principles of authentic OSCE station design [25].OSCE stations vary considerably in their design and focus and this station cannot be assumed to represent all OSCE stations. Despite this, similar stations are fairly frequent which may aid extrapolation. The station was scored by participants using the GeCos scoring system [26,27]. This involved participants scoring candidates on five relevant domains: (history content; history process; clinical reasoning; management content; building and maintaining the relationship) on a scale of 1–4 (1 = must improve; 2 = borderline; 3 = proficient; 4 = very good), and a further 7-point global scale (1 = incompetent; 7 = excellent). As per the intent of the scale [26], these domains were summed to give a score out of 27. Participants were asked to view a scoring orientation video prior to workshop attendance to facilitate familiarity with the format.

Developing benchmark information and agreed scores. The benchmark information, which formed part of the intervention, was generated for 2 separate performances, which were selected to represent overall levels of performance which were Good and Borderline. We chose this combination of videos as Edwards et al.'s Realist evaluation of VBB [19] specifically suggested that this combination may offer a suitable compromise between the burden on examiners and the range of exemplified performance. Benchmark information comprised of 1/a pre-defined 'agreed' score (i.e. the consensus total score reached by expert panels), 2/a brief written justification for that score. The scores were generated through iterative review by two expert examiner panels. The first panel, consisting of four examiners from the lead institution, screened videos for suitability as study materials. The second panel, consisted of 11 examiners drawn from 9 UK medical schools, including 2 from the lead institution. All members were general practitioners or consultant doctors and either OSCE leads, assessment leads or senior external examiners. As the station content was considered generic to medicine, no specific content expertise was required. Panel members were excluded from further participation in the study. Panel members independently assessed and provided feedback for a narrower selection of videos and then discussed the reasoning for their scores with other panel members with the aim of reaching consensus. Panel members then provided adjusted scores and feedback. The mean of their adjusted total scores for each video were taken as the final benchmark scores. Target scores for these performances were based on the overall score allocated to the video (out of 27 on the GeCoS scale) whilst classification was performed using the 7-point global ratings component of the GeCoS scale. For the two benchmark videos (i.e. for use in the intervention), the final panel-allocated score for the borderline performance was 16 out of 27 (range: 13–18) with a global score of 3 (Unsatisfactory) and for the excellent performance was 24 out of 27 (range: 20–26) with a global score of 6 (Very Good). For the two target videos (used to compare the scoring of both groups), the final panel-allocated score for the borderline video was 15 out of 27 (range: 10–21) with a global score of 3 (Unsatisfactory) and for the good video was 20 out of 27 (range: 15–25) with a global score of 5 (Good). The written justifications were based on the feedback provided by members of the expert panel; the panel feedback was reviewed by the research team and summarised to include the key performance aspects.

Procedure

Participants received the station information and scoring orientation video up to one week before their scheduled workshop. Station information described the station scenario, simulated patient script, the scoring format and rubric and gave brief written guidance on the examiner role. The scoring orientation training explained the scoring format and then asked examiners to practise using it on a video-performance which was unrelated to the station in the study. Examiners were asked to use these materials to prepare, but to ensure ecological validity with usual practice, this was not enforced and was not monitored. As a result, examiners may have varied in their degree of preparation.

Participants attended either a control or intervention workshop, depending on their allocation, in groups of 3–10 *via* Microsoft Teams. All participants viewed videos during the workshop and took part in the group discussion.

The workshop procedure differed depending on experimental group allocation, illustrated by Figure 1. In the intervention, participants first assessed two benchmark performances (excellent and borderline), each followed by the associated benchmark information. Then, the group had a guided 10-minute discussion of the materials to resolve discrepancies through consultation with their peers. The benchmarking

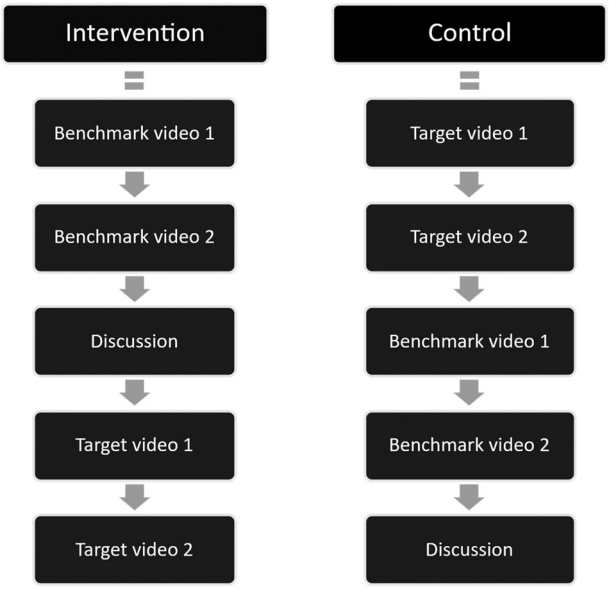


Figure 1. Study procedure by experimental group allocation.

procedure lasted approximately 35 min. Next, participants provided a 'confidence' rating in their ability to rate subsequent performances on that station. Confidence ratings asked the question 'How likely do you think you are to score the following performances in a way that aligns with the average of a wider group of examiners scoring the same performances?'. This was answered on a 7-point end-anchored Likert scale (1- Not likely at all; 7- Very likely). Confidence ratings were presented as 'perceived alignment' as the use of the word 'confident' has been shown to impact cognitive performance [28]. Moreover, we phrased this as 'alignment with a wider group' rather than 'alignment with the expert panel' because the control group (who answered the same question) could not reasonably be asked to judge alignment with an expert panel whose judgement they had never encountered.

Finally, participants assessed two target performances (borderline and good), the order of which was counterbalanced between workshops to avoid order effects.

The control group first provided a confidence rating (same format as intervention group) and then assessed the two target performances, again counterbalanced between workshops. Next, in order to balance the workload across intervention and control groups, participants viewed the two benchmark performances and associated benchmark information, followed by the guided discussion.

Analysis

For each question, we used either Generalised Linear Mixed-effects Models (GLMM) or Generalized Linear Models (GLM) [29] with theoretically derived pre-specified model terms. This resulted in 6 similar yet distinct statistical models. Models included all expected sources of variance (performance, group allocation, school, order) with specific relevant interactions as described. Depending on the data type and structure, different family and link functions were used to ensure adequate model fit. All analyses were conducted in R version 4.4.1 [30] using lme4 [31], stats [30,31] and glmmTMB [32], depending on model type and data requirements. Details of how fixed and random effects were determined for each analysis can be found in appendix 2.

Consistent with CONSORT guidelines, no inferential statistics were calculated on participant baseline characteristics [33]. Pre-specified outcome variables were used for each research question. Examiner score variability (RQ1 &2) was calculated as the absolute difference from the group mean for each video viewed ($|x - \bar{x}|$, where x is the participants overall score and $\bar{x}$ is the group mean of the overall score for the relevant target video). Overall score was calculated as the sum of the domain and global scores from the GeCoS scale, with a possible range of 6–27 points. *Pass-fail category* (RQ3) was determined by whether the participants' global score descriptor fell on the pass (points 4–7) or fail (1–3) side of the scale for each video. *Accuracy score* was calculated as the difference between the participants score and expert panel score for each video, using the same method as score variability. No interim analyses were performed. All analyses were performed on all randomised participants except for RQ 2 (Does VBB reduce variability in outlying examiners), where 'outliers' were identified based on the criteria that their scores were more than 1.5 standard deviations from 0 on the variability scale (with 0 indicating no variability from the group mean) for their first viewed video (i.e. first video variability $>$ 3.038). The criterion of 1.5 standard deviations was selected *a priori* to

represent a pragmatic cut-off on the basis that it was judged likely to produce a sufficient subset for analysis whilst representing examiners who had scored differently from the majority. It was determined solely from the first-viewed video to avoid influence of benchmarking material on the classification.

Results

407 participants were recruited. Of these, 244 consented to participate and were scheduled for workshops. Thirty-seven workshops were run (20 intervention, 17 control), which were conducted between November 2023 and June 2024. Dropouts were roughly even by group (Intervention 39; Control 34). In total, 171 participants from 12 medical schools completed the study. All participants who completed the study received either the intervention or control as per protocol. Complete data were available for all participants who were included in all analyses except the 'outliers' subset analysis where just those identified by the described method were included. The trial was stopped because participation exceeded the sampling target. Participation was balanced across intervention ($n = 87$) and control ($n = 84$) (see: Table 1 for demographics, and Appendix 1 for Consort diagram). Participation varied by school (range 2 to 24), but no substantial imbalances in group allocation by school occurred.

Overall scores for performances were normally distributed. The benchmark videos received mean scores of 16.49 (95% CIs 15.91–17.07) for the borderline performance and 24.09(95% CIs 23.64–24.54) for the good performance and hence were very similar to the expert panel's scores. Target videos received means scores of 18.31 (95% CIs 17.67–18.95) for the borderline performance and 20.11(95% CIs 19.51–20.71) for the good performance. As a result, the borderline video received a slightly higher mean score than the expert panels agreed score.RQ1: Does VBB reduce overall examiner variability?

Examiner variability data showed a positively-skewed distribution (as expected, given that it was based on absolute difference from the group mean), so were analysed using a Gamma log link GLMM. The random effect of school was removed as it did not contribute to the variance in the model, resulting in singular fit.

$$Variability \sim Allocation + Video + Allocation * Video + School + Order + (1 \mid Participant)$$

Median examiner variability in the intervention group was 3.02 (1st quartile 1.46, 3rd quartile 4.78) and the control group was 2.51 (1st quartile 1.30, 3rd quartile 4.70) (see Figure 2), model parameter estimates −0.053 (95% CIs −0.148, 0.042) p value = 0.273, thereby indicating no significant difference.

Table 1. Participant demographics by group.

Demographic	Category	Control $n = 84$ (% of group)	Intervention $n = 87$ (% of group)	Total
Clinical grade	Career grade	8(9.52%)	5(5.75%)	13
	Consultant	34(40.48%)	34(39.08%)	68
	General practice	22(26.19%)	27(31.03%)	49
	Resident	18(21.43%)	20(22.99%)	38
	Unknown	2(2.38%)	1(1.15%)	3
Specialty	Anaes/ICU	4(4.76%)	9(10.34%)	13
	Child health	7(8.33%)	8(9.20%)	15
	Diagnostic	2(2.38%)	6(6.90%)	8
	Emergency medicine	9(10.71%)	0(0.00%)	9
	Medicine	19(22.62%)	10(11.49%)	29
	O&G	2(2.38%)	6(6.90%)	8
	Other	5(5.95%)	2(2.30%)	7
	General practice	26(30.95%)	33(37.93%)	59
	Psychiatry	3(3.57%)	5(5.75%)	8
	Surgery	7(8.33%)	8(9.20%)	15
OSCE examiner experience	0–2 years	27(32.14%)	27(31.03%)	54
	3–5 years	26(30.95%)	24(27.59%)	50
	6–9 years	7(8.33%)	8(9.20%)	15
	10+ years	24(28.57%)	28(32.18%)	52
School	A	7(8.33%)	7(8.05%)	14
	B	7(8.33%)	7(8.05%)	14
	C	11(13.10%)	8(9.20%)	19
	D	5(5.95%)	6(6.90%)	11
	E	5(5.95%)	8(9.20%)	13
	F	10(11.90%)	9(10.34%)	19
	G	8(9.52%)	10(11.49%)	18
	H	8(9.52%)	9(10.34%)	17
	I	0(0.00%)	2(2.30%)	2
	J	13(15.48%)	12(13.79%)	25
	K	5(5.95%)	6(6.90%)	11
	L	5(5.95%)	3(3.45%)	8

Please note identifying letters for schools were randomly generated but are used consistently within presented data. Where clinical grade is unknown, participants opted not to supply this information.

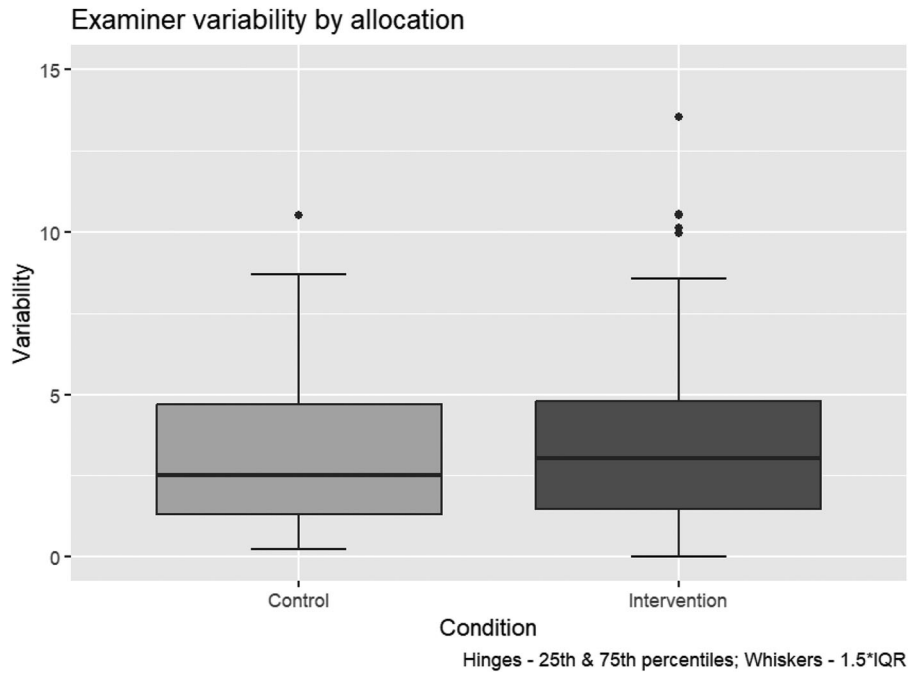


Figure 2. Boxplot of examiner variability by control and intervention groups across target videos. Data represents median variability, inter-quartiles ranges and full range of data. Units are in the same scale as the original assessment.

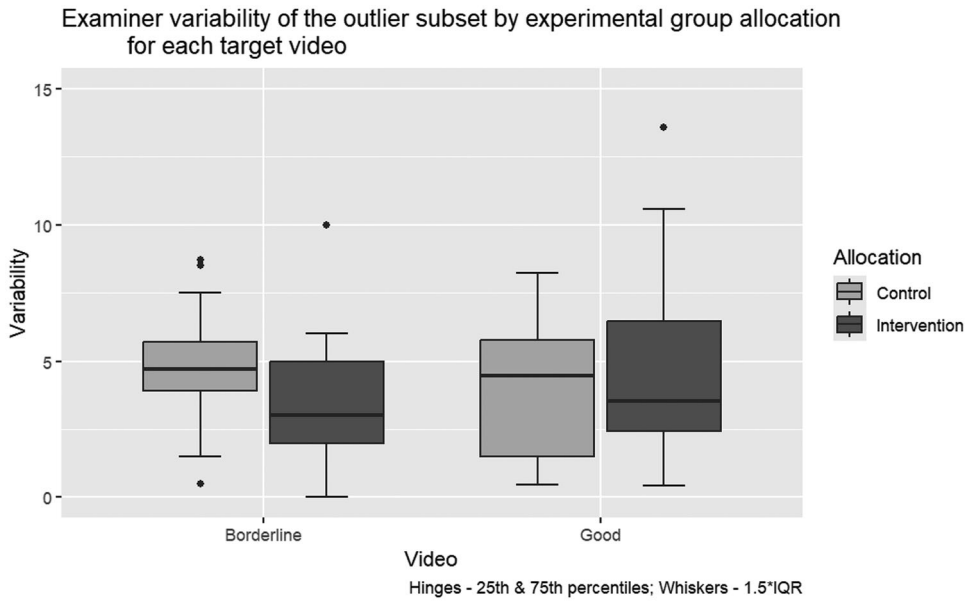


Figure 3. Boxplot of examiner variability of the outlier subset by experimental group allocation for each target video.

Main effects of performance (0.070 (−0.018, 0.158), $p = 0.119$), school (see Table s2, appendix 2 for estimates, $p = 0.067$–0.930) and order (0.028 (−0.072, 0.128), $p = 0.587$) were not statistically significant and there was no statistically significant interaction for group (intervention vs control) x performance (good vs borderline) (0.057 (−0.03, 0.145), $p = 0.198$).

RQ2: Does VBB reduce examiner variability for outlying examiners?

Data showed a slight positive skew, so a Gaussian log link GLM was used to analyse the data. Random effects were removed as they did not contribute to the variance in the model, resulting in singular fit.

$$\begin{aligned} \textit{Outlier Variability} \sim\ & \textit{Allocation} + \textit{Video} \\ & + \textit{Allocation} * \textit{Video} \\ & + \textit{School} + \textit{Order} \end{aligned}$$

60 examiners were in the subset of participants considered 'outliers' of whom 25 were in the intervention and 35 were in the control group. No significant effects were found for order or video. Within this outlier subset of examiners, school J had higher median examiner variability (5.78 (1st quartile 3.39, 3rd quartile 6.70) $p < 0.03$) than the reference

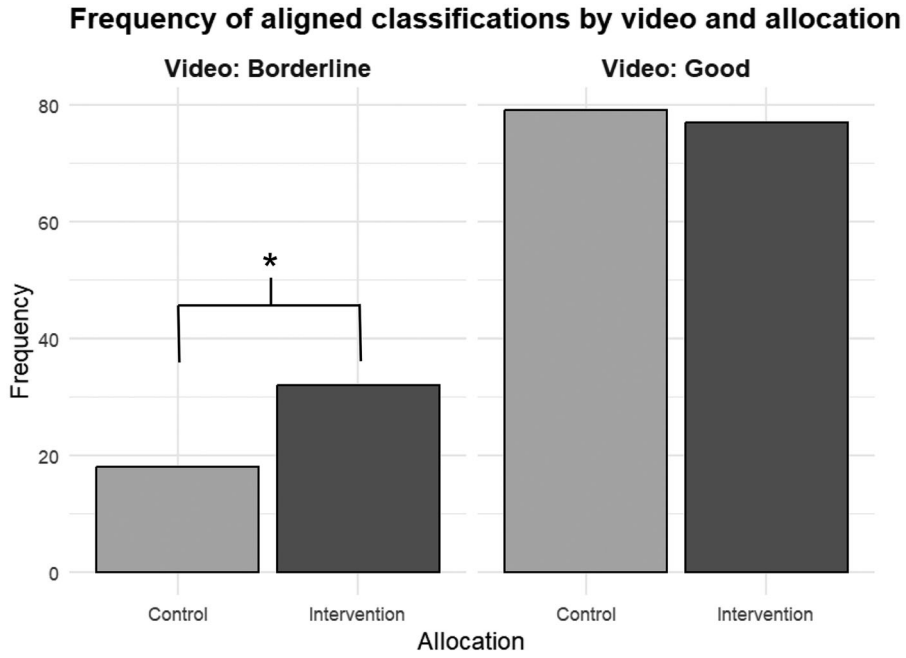


Figure 4. Frequency of aligned classifications by video and experimental group allocation.
*Indicates statistical significance, $p < 0.05$

standard (4.88 (4.67–5.09). No other schools were statistically significantly different. Median examiner variability in the intervention group was 3.45 (1st quartile 2.02, 3rd quartile 5.53) and the control group was 4.60 (1st quartile 2.69, 3rd quartile 5.76), estimate = 0.000 (−0.115, 0.123), $p = 0.997$, thereby indicating no significant difference. In contrast to the first RQ (overall examiner variability), there was a significant interaction of group (intervention vs control) x performance (good vs borderline), estimate = 0.143(0.030, 0.266), $p = 0.016$, indicating that there was a difference in effects across performances. Inspection of parameters for individual performances showed that for the borderline performance intervention group had median variability of 3.02 (1st quartile 1.98, 3rd quartile 4.98) vs control group median variability of 4.70 (1st quartile 3.91, 3rd quartile 5.70), whereas for the good video intervention group median variability was 3.54 (1st quartile 2.43, 3rd quartile 6.46) vs control group median variability of 4.47 (1st quartile 1.50, 3rd quartile 5.78) (see Figure 3). Consequently, in this subset of outlying examiners, the intervention group showed less variability than the control group in their scores for the borderline performance.

RQ3: Does VBB increase the proportion of correctly classified performances?

Correct classification showed a binary distribution, as expected for the binary nature of this outcome variable (correct vs incorrect). Therefore, the data were analysed with a binomial family GLM and logit link function. Random effects were removed as they did not contribute to the variance in the model, resulting in singular fit.

$$\begin{aligned} \textit{Correct classification} \sim\ & \textit{Allocation} + \textit{Video} \\ & + \textit{Allocation} * \textit{Video} \\ & + \textit{Order} \end{aligned}$$

A correct response was considered as a classification of the target performance on the global scale that aligned with the expert panel's classification of the performance; an incorrect response was misaligned with the expert panel classification.

In the control group 57.74% of participants correctly classified performances compared to 62.64% in the intervention group. No significant main effect of group (model estimate = −0.007 (−0.329,0.336), $p = 0.966$) nor order (estimate = 0.241 (−0.044, 0.533), $p = 0.100$) was found.

There was a significant main effect of performance level with 29.24% of participants classifying the borderline performance correctly versus 91.23% classifying the good video correctly (estimate = −1.679 (−2.032, −1.365), $p < 0.001$). This indicates that the good performance was classified correctly much more frequently than the borderline performance.

A significant interaction of video by group was found (estimate = −0.372 (−0.716, −0.052), $p < 0.026$), illustrated by Figure 4. 21.43% of participants in the control group classified the borderline video correctly versus 36.78% in the intervention group, with contrasts showing a significant difference of the rate of correct classifications for the borderline video between the control and intervention groups ($p < 0.030$). This gave an odds ratio of correctly classifying the borderline performance of OR = 2.133 (95% CI 1.081–4.208). For the good video, 94.05% in

the control group made the correct classification, whereas, in the intervention group, 88.51% made the correct classification – this difference was not significant ($p = 0.202$).

RQ 4: Does VBB increase accuracy (i.e. distance from agreed 'correct' score)?

Accuracy scores were fitted with a Gaussian family GLM with an identity link function, as the data were approximately normal. Random effects were removed as they did not contribute to the variance in the model, resulting in singular fit.

$$Accuracy \sim Allocation + Video + Allocation * Video + Order$$

Median accuracy in the intervention group was 3.02 (1st quartile 1.46, 3rd quartile 4.78) and the control group was 2.51 (1st quartile 1.30, 3rd quartile 4.70), estimate = −0.082(−0.521, 0.357), p- = 0.714, thereby indicating no significant difference. No significant effects of order or the interaction of experimental group and video were found. A significant effect of performance was found, estimate= 1.609 (1.170, 2.048), $p < 0.001$, indicating that examiners were more accurate in their ratings for the good performance (median 0.00, 1st quartile −2.50, 3rd quartile 3.00) than the borderline performance (median 3.00, 1st quartile 0.00, 3rd quartile 6.5).

RQ5: Does VBB increase judgmental confidence?

Judgemental confidence ratings showed a left skew; therefore, the data were modelled using a Gamma family log link GLMM. School was included as a random effect. The data also showed issues with homogeneity of variance, so the model was fitted with a dispersion parameter to account for heteroscedacity (see appendix 2 for further details). A random effect for participants was not included, as there was only one observation per participant.

$$Full\ model: Alignment \sim Allocation + (1 \mid School)$$
$$dispformula = \sim Allocation + School$$

Median confidence for the control group was 5.00 (1st quartile 4.75, 3rd quartile 5.25) and the intervention group was 5.00 (1st quartile 5.00, 3rd quartile 6.00), estimate = 0.049 (0.021, 0.077), $p < 0.001$. Therefore, despite the median values being the same between groups, this analysis indicated a small, statistically significant difference in confidence between control and intervention groups.

RQ6: Are inter-school differences apparent in examiner scoring?

Overall score showed an approximately normal distribution; therefore, the data were modelled using a Gaussian family identity link GLMM. Participants were included as a random effect.

$$Overall\ Score \sim School + Allocation * Video + Order + (1 \mid Participant)$$

Median overall scores by school for the borderline target video ranged from 15.00–19.00; and 17.00–24.00 for the good target video.

No significant effects of school for overall scores were found ($p = 0.469$–0.900, see Figure s1 and Table s7 (appendix 2) for parameter estimates). No main effect of order was found ($p = 0.881$); no significant interaction was found in overall scores for group (intervention vs control) x performance (good vs borderline) (estimate= 0.373 (−0.054, 0.800), $p = 0.086$).

As RQ 7 was dependant on a positive result for RQ6, it was not addressed. Full details of all model parameter estimates can be found in appendix 2.

Discussion

Summary of results

This study used a prospectively-powered, open-label, randomised, stratified, controlled trial to compare the influence of video-based benchmarking with control on examiner variability within a simulated OSCE context. It sampled nationally, exceeding its pre-specified recruitment target of 160 examiners, including a total of 171 examiners drawn from 12 UK medical schools. It did not show a reduction in overall examiner variability (its primary end-point) but has shown an effect on two of its pre-specified secondary endpoints, namely a/examiner variability for 'outlying' examiners and b/correct classification of performance, although both of these findings were limited to the borderline performance (they were not present for the good performance). VBB had no influence on accuracy of examiners' scoring, but it showed a statistically significant, but small influence on examiners' confidence. In contrast to some prior work [23], it did not show a difference in the average standard of scoring between examiners from different medical schools.

Relationship to prior literature and theoretical interpretation

Video-based benchmarking as an intervention is grounded in two theoretically derived propositions, namely that to influence examiners frame of reference sufficiently to align their scoring, training examples needs to be *close in time* and *specific* to the examination task. Despite using both of these conditions and incorporating further recommendations

from recent work by Edwards et al.'s [20] (a/use of multiple benchmarking videos showing different levels of performance and b/the opportunity for examiners to discuss with peers), we found no overall influence of VBB on examiners' variability. In this regard our findings align with prior research [11,12]. This study complements earlier realist work on VBB [20] by testing its measurable impact at scale. While realist approaches explored how contextual and individual factors might shape examiner alignment, our objectivist design provides evidence about whether such alignment is reliably achieved. As described earlier, Kogan et al. [13] showed a moderate effect of more prolonged training which involved longitudinal engagement with examiners through an intensive series of in-person workshops at the start and then shorter follow up on-line workshops over a 3-month period. These observations suggest that (in contrast to our theoretical premise) the total amount of training may be more important to aligning examiners' judgements than either specificity of training material or proximity of the training in relation to testing. A range of reasons may account for this: examiners in our study could (for example) have had insufficient time to reflect on the benchmarking information in order to achieve its maximal benefits. Some participants in Edwards et al.'s study commented that whilst they felt that any influence of benchmarking would atrophy over time, having it too close to the OSCE could potentially lessen its effect [19]. Equally, it may be that more examples, spread out over time (i.e. using the cognitive learning strategy of spaced repetition34] accounts for the greater effect which Kogan and colleagues achieved.

Whilst we did not demonstrate an overall effect on examiner variability, we saw effects on two secondary endpoints in relation to the borderline video. Given that the purpose of graduation-level OSCE exams is typically to establish mastery [35], decisions about performances close to the pass/fail boundary have particular importance for the outcome of the assessment, especially in settings where conjunctive passing rules are used [36] as individual station results can have an important bearing on overall outcomes. Leniency or 'giving the benefit of the doubt' [37] at this level could potentially pose a challenge to patient safety as it may enable candidates who haven't met the required standard to progress, thereby posing a challenge to the 'consequences' component of the assessment's validity [7]. Examiners describe greater uncertainty around borderline performances than they do for either good or clearly failing performances [38]. Whilst they recognise that the performance is weak, they struggle to determine whether it should pass or fail. In our study the majority of examiners passed the borderline (weakly failing) performance. Whilst it is possible to debate whether the expert panel were 'correct' to determine it was a failing performance, our data indicate that regardless of this, video-based benchmarking made it more than twice as likely that examiners would categorise the performance in-line with the standard which the panel set. It therefore clearly had an influence of bringing examiners judgements in line with the expert panel, suggesting that VBB has utility in clarifying the threshold which examiners should apply. VBB also reduced the observed variability for outlying examiners. This is also important. Sampling across stations and examiners is the principal means by which OSCEs achieve reliability [2,39], however, owing to central limit theorem [40], outlying examiners have a disproportionate influence on the mean of a score. Therefore, it is of particular interest that VBB may have an influence on this subgroup of examiners and is consistent with prior suggestions that calibration interventions can help to align outlying examiners scoring [41]. For all of these reasons, these positive secondary end-points may indicate that despite missing its primary endpoint, VBB still has beneficial effects.

Implications for practice

Our findings suggest that while VBB may not reduce overall examiner variability, it may offer specific benefits in improving borderline scoring and reducing outlier behaviour. Educators might consider VBB as a targeted calibration tool for stations with known variability or critical pass/fail thresholds. VBB uses station-specific benchmarking performances. Consequently, the resource implications of generating benchmark material for all OSCE stations may be considerable. In this study, VBB was implemented collaboratively across multiple medical schools, using shared benchmark videos and structured discussion to promote a common interpretive frame. This approach may be especially valuable in national or inter-institutional assessments where consistency across sites is a concern. Equally, benchmarking material from a smaller number of typical stations may be sufficient. This would require empirical exploration. VBB imposes a demand on examiners' time. Examiners in prior research [19] considered that performing the benchmark intervention within 24-hours of the OSCE would likely be effective. This may help to manage this time burden. Equally, rather than recalibrating all examiners, VBB may be most effective when used selectively with new, infrequent, or previously inconsistent examiners. Its value lies not only in improving judgment, but in building shared understanding across institutions.

Limitations

Our study had many strengths, including its stratified randomised controlled design, broad participant sampling, and use of authentic stimuli. Despite this, it has some limitations. Our stimulus material only included one clinical scenario; we can't exclude the possibility that different findings might occur with different case material. Our VBB intervention, based on prior research, used a good and poor video. We can't exclude the potential that a different combination of videos, perhaps including poor performances, could have achieved different results. Whilst our study was prospectively powered to find a 30% reduction in examiner variance, and set conservative recruitment targets which were exceeded, we can't exclude the possibility that a larger sample size could have detected smaller differences in our non-significant findings. Our study was conducted in the context of undergraduate medical education. Whilst we expect considerable transferability to other settings, the findings don't implicitly generalise. Whilst it was necessary to use a simulated setting to enable controlled comparisons, we can't exclude the possibility that other effects could occur in naturalistic settings. Similarly, volunteer examiners may perform differently than examiners in real settings. Findings regarding outlying examiners were based on subset analysis and should ideally be prospectively reified *via* a method that prescreens to find 'outliers' and then randomises these individuals to either VBB or control.

Recommendations for future research

Future research should seek to replicate these findings across different settings and a mix of OSCE stations and explore VBB use within real OSCEs. Follow up work which specifically randomises 'outlying' examiners may help to verify our findings.

Conclusions

Whilst VBB demonstrated no overall effect on examiner variability, our findings suggest it may help examiners to correctly classify borderline performances and may help to align examiners who are further from the group mean as well as producing small benefits to examiners' confidence. Collectively these effects suggest that VBB may benefit OSCE validity and enhance trust in an assessment. As engagement by examiners with VBB is critical, institutions should carefully construct benchmark material which may therefore be well served through assessment collaborations.

Acknowledgements

The authors would like to thank all the examiners who participated in the study and the students who allowed their videos to be used as stimulus material. We would like to thank the Medical Schools Council Assessment Alliance (MSCAA) for facilitating centre recruitment.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was funded by an unrestricted grant from the MPS Foundation. The opinions expressed in this paper represent the opinions of the authors and are not necessarily the position of the MPS Foundation.

Notes on contributors

Peter Yeates, Keele University.

Rebecca Jane Edwards, University of Leicester.

Aditya Narain, Keele University.

Robert McKinley, Keele University.

Janet Lefroy, Keele University.

Gareth McCray, Keele University.

Giles Roberts, Keele University.

Ellie Hammond, Keele University.

Stu McBain, Keele University.

Andrew Blythe, University of Bristol.

Kathy Cullen, Queen's University Belfast.

Craig Napier, University of Glasgow.

Laura Sims, University of Exeter.

Harish Thampy, University of Manchester.

Tushar Vince, King's College London.

Sue Ensaff, Cardiff University.

Rhian Goodfellow, Cardiff University.

Christopher Harrison, University of Central Lancashire.

Ching-Wa Chung, University of Aberdeen.

Steven Capey, Swansea University.

Chris Roberts, University of Sheffield.

Rebecca Vallender, Cardiff University.

ORCID

Peter Yeates http://orcid.org/0000-0001-6316-4051
Aditya Narain http://orcid.org/0000-0003-3947-7925
Robert McKinley http://orcid.org/0000-0002-3684-3435
Gareth McCray http://orcid.org/0000-0002-0728-5171
Andrew Blythe http://orcid.org/0009-0000-2473-7996
Kathy Cullen http://orcid.org/0000-0001-6207-0491
Harish Thampy http://orcid.org/0000-0002-7850-4378

Data availability statement

The anonymous study dataset is available from the authors on the basis of reasonable request.

Video-based Benchmarking

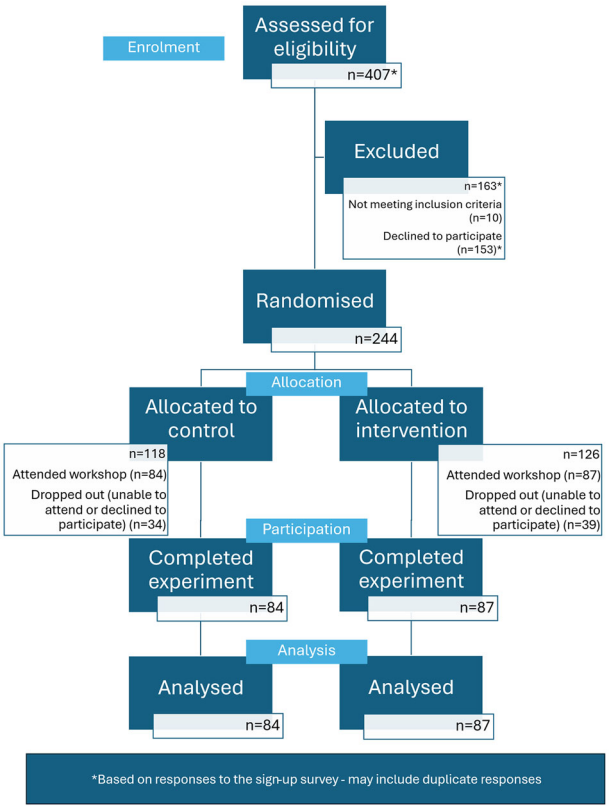
An intervention which aims to align examiners' judgements in OSCEs with a pre-agreed standard. It involves examiners watching and scoring one or more station-specific videos of candidate performances which are specific to the station they will examination, shortly before the OSCE occurs. Examiners then compare their scores to the agreed scores, consider justifications for the agreed scores and reflect on and discuss any differences.

References

1. Epstein RM, Hundert EM. Defining and assessing professional competence. J Am Med Assoc. 2002;287(2):226–235. doi: 10.1001/jama.287.2.226
2. Harden R, Lilley P, Patricio M. The definitive guide to the OSCE. The Objective Structured Clinical Exam as a Performance Assessment. Elsevier Health Sciences; 2015.
3. Boursicot K, Kemp S, Wilkinson T, et al. Performance assessment: consensus statement and recommendations from the 2020 Ottawa Conference. Med Teach. 2021;43(1):58–67. doi: 10.1080/0142159X.2020.1830052
4. Tavares W, Kuper A, Kulasegaram K, et al. The compatibility principle: on philosophies in the assessment of clinical competence. Adv Health Sci Educ Theory Pract. 2020;25(4):1003–1018. doi: 10.1007/s10459-019-09939-9
5. Tighe J, McManus IC, Dewhurst NG, et al. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. BMC Med Educ. 2010;10(1):40. doi: 10.1186/1472-6920-10-40
6. Kogan JR, Conforti LN, Iobst WF, et al. Reconceptualizing variable rater assessments as both an educational and clinical care problem. Acad Med. 2014;89(5):721–727. doi: 10.1097/ACM.0000000000000221
7. Cook DA, Brydges R, Ginsburg S, et al. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49(6):560–575. doi: 10.1111/medu.12678
8. 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. doi: 10.1080/17437270802124525
9. Pell G, Fuller R, Homer M, et al. How to measure the quality of the OSCE: a review of metrics - AMEE guide no. 49. Med Teach. 2010;32(10):802–811. doi: 10.3109/0142159X.2010.507716
10. Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. Med Teach. 2018;40(12):1208–1213. doi: 10.1080/0142159X.2017.1390214
11. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence. Ann Intern Med. 2004;140(11):874–881. MD MD doi: 10.7326/0003-4819-140-11-200406010-00008
12. Cook DA, Dupras DM, Beckman TJ, et al. 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. doi: 10.1007/s11606-008-0842-3
13. Kogan JR, Dine CJ, Conforti LN, et al. Can rater training improve the quality and accuracy of workplace-based assessment narrative comments and entrustment ratings? A randomized controlled trial. Acad Med. 2023;98(2):237–247. doi: 10.1097/ACM.0000000000004819
14. Yeates P, O'Neill P, Mann K, et al. 'You're certainly relatively competent': assessor bias due to recent experiences. Med Educ. 2013;47(9):910–922. doi: 10.1111/medu.12254
15. Yeates P, Cardell J, Byrne G, et al. Relatively speaking: contrast effects influence assessors' scores and narrative feedback. Med Educ. 2015;49(9):909–919. doi: 10.1111/medu.12777
16. Murto SH, Shaw T, Touchie C, et al. Are raters influenced by prior information about a learner? A review of assimilation and contrast effects in assessment. Adv Health Sci Educ. 2021;26(3):1133–1156. doi: 10.1007/s10459-021-10032-3
17. Shaw T, Wood TJ, Touchie C, et al. How biased are you? The effect of prior performance information on attending physician ratings and implications for learner handover. Adv Health Sci Educ Theory Pract. 2021;26(1):199–214. doi: 10.1007/s10459-020-09979-6
18. Uggerslev KL, Sulsky LM. Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. J Appl Psychol. 2008;93(3):711–719. doi: 10.1037/0021-9010.93.3.711
19. Edwards RJ, Yeates P, Lefroy J, et al. Understanding contexts and mechanisms through which video based benchmarking promotes alignment of examiners' scoring in objective structured clinical exams. Adv Health Sci Educ Theory Pract. Published online July 4, 2025. doi: 10.1007/s10459-025-10454-3
20. Edwards R, Yeates P, Lefroy J, et al. Video-Based Benchmarking: an intervention to increase OSCE examiner alignment. In Ottawa Conference on Assessment in Health Professionals Education. Assessment Near and Far; 2024.
21. Tweed M, Ingham C. Observed consultation: confidence and accuracy of assessors. Adv Health Sci Educ Theory Pract. 2010;15(1):31–43. doi: 10.1007/s10459-009-9163-5
22. Heaslip V, Scammell JME. Failing underperforming students: the role of grading in practice assessment. Nurse Educ Pract. 2012;12(2):95–100. doi: 10.1016/j.nepr.2011.08.003
23. Yeates P, Maluf A, McCray G, et al. Inter-school variations in the standard of examiners' graduation-level OSCE judgements. Med Teach. Published online July 8. 2025;47(4):735–743. 1-las. doi: 10.1080/0142159X.2024.2372087
24. Yeates P, Edwards R. Experimental methods: more than just testing a hypothesis. In: E Rees, A Ledger, K Walker, editors. Starting research in clinical education. Wiley Blackwell; 2024. p. 191–199.
25. Yeates P, Maluf A, Kinston R, et al. A realist evaluation of how, why and when objective structured clinical exams (OSCEs) are experienced as an authentic assessment of

clinical preparedness. Med Teach. 2025;47(3):458–466. doi: 10.1080/0142159X.2024.2339413
26. Lefroy J. GeCoS Generic Consultation Skills toolkit. 2024 [cited 2024 Nov 13]. Available from: https://www.keele.ac.uk/gecos/
27. Lefroy J, Gay SP, Gibson S, et al. Development and face validation of an instrument to assess and improve clinical consultation skills. Int J Clin Skills. 2011;5(2):115–125.
28. Double KS, Birney DP. Are you sure about that? Eliciting confidence ratings may influence performance on Raven's progressive matrices. Think Reason. 2017;23(2):190–206. doi: 10.1080/13546783.2017.1289121
29. Field A. Discovering statistics using SPSS. 3rd ed. London: Sage; 2009.
30. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Preprint posted online 2021. Available from: https://www.r-project.org/.
31. Bates D, Maechler M, Bolker B, et al. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67(1):1–48. doi: 10.18637/jss.v067.i01
32. Brooks ME, Kristensen K, Benthem KJ, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. 2017;9(2):378. van. doi: 10.32614/RJ-2017-066
33. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012;10(1):28–55. doi: 10.1016/j.ijsu.2011.10.001
34. Weinstein Y, Madan CR, Sumeracki MA. Teaching the science of learning. Cogn Res Princ Implic. 2018;3(1):2. doi: 10.1186/s41235-017-0087-y
35. GMC. Requirements for the MLA Clinical and Professional Skills Assessment [cited 2023 Dec 20]. Available from: https://www.gmc-uk.org/education/medical-licensing-assessment/uk-medical-schools-guide-to-the-mla/clinical-and-professional-skills-assessment-cpsa
36. Homer M, Russell J. Conjunctive standards in OSCEs: the why and the how of number of stations passed criteria. Med Teach. 2021;43(4):448–455. doi: 10.1080/0142159X.2020.1856353
37. Wong WY. Consistency of examiner judgements in medical education: a case study of competency-based assessment. The University of Queensland; 2018.
38. Wong WYA, Thistlethwaite J, Moni K, et al. Using cultural historical activity theory to reflect on the sociocultural complexities in OSCE examiners' judgements. Adv Health Sci Educ Theory Pract. 2023;28(1):27–46. doi: 10.1007/s10459-022-10139-1
39. 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. doi: 10.1016/j.jarmac.2018.03.008
40. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol. 2017;70(2):144–156. doi: 10.4097/kjae.2017.70.2.144
41. Sturman N, Ostini R, Wong WY, et al. 'On the same page'? The effect of GP examiner feedback on differences in rating severity in clinical assessments: a pre/post intervention study. BMC Med Educ. 2017;17(1):101. doi: 10.1186/s12909-017-0929-9

Appendix 1:
CONSORT diagram



Appendix 2:
Statistical appendix
Process of selection of fixed or random intercepts for variables

Random intercepts for School and Participant (nested within School) were included to account for clustering of observations within participants and schools. Random slopes for Allocation were initially specified in selected models to allow the effect of Allocation to vary across groups, accounting for potential between-school or between-participant heterogeneity, particularly given that some schools had relatively small sample sizes.

Where models produced singular fits, variance components of the random effects were examined. Model complexity was reduced by first removing the random slope for Allocation, followed by removal of random-effect terms with variance estimates close to zero, assessed *via* diagnostic plots.

Allocation was initially included as a random slope in the models of overall variability (RQ1) and classification accuracy (RQ3). These models consistently resulted in singular fits and indicated negligible variance associated with the Allocation random slope. Because subsequent outcome measures were conceptually similar (variability and accuracy metrics), and Allocation was retained as a fixed effect in all models, the random slope for Allocation was not included in later analyses.

For the confidence outcome (RQ5), only a random intercept for School was included. Allocation was modelled as a fixed effect, and each participant contributed only a single observation. As random slopes require repeated observations within grouping units, a Participant-level random effect was not specified for this model.

Explanation of dispersion parameter to account for heteroscedacity

A dispersion parameter was included to account for heteroscedasticity in the Alignment outcome. Diagnostic checks of simpler Gaussian and Gamma mixed models revealed non-constant residual variance and patterns in residual plots, indicating violation of distributional assumptions.

In glmmTMB, the dispersion model allows the residual variance to vary as a function of predictors (here, Allocation and School), rather than assuming constant variance across observations. Including a dispersion structure improved residual diagnostics and model fit (AIC/BIC).

The dispersion parameter therefore ensured more robust estimation of standard errors and improve validity of statistical inference.

Random effects were retained in analyses for RQs 1,5 and 6. Random effects in other models produced singular fit and so were dropped.

* Residual variance and ICC are not reported because a dispersion parameter was fitted, allowing residual variance to vary by Allocation and School. As a result, residual variance is not constant across clusters and a single ICC cannot be meaningfully calculated.

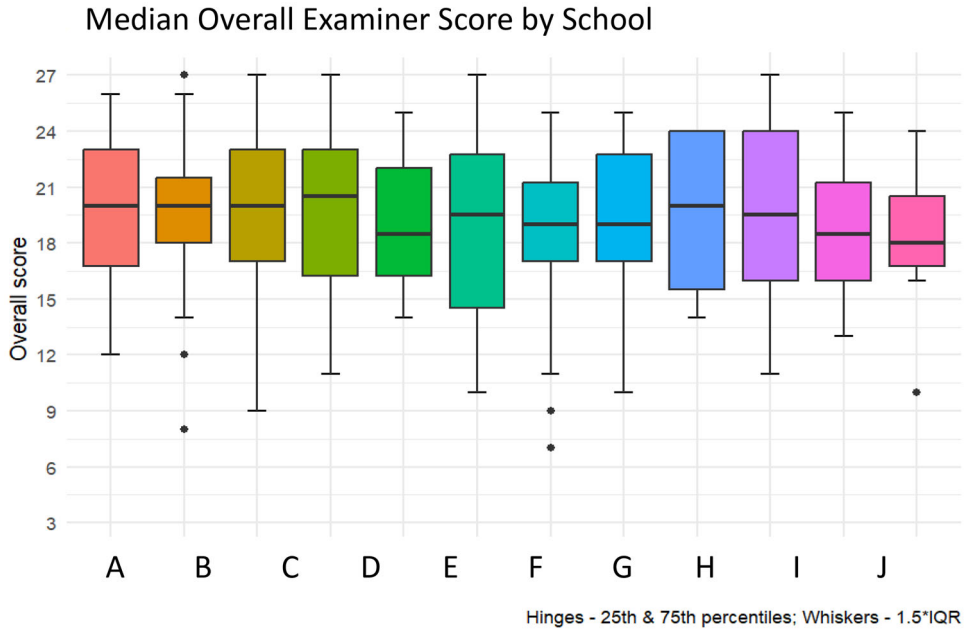


Figure A1. Boxplot of Median Overall Examiner Score by School.

Table A1. Variance attributable to random effects in each model where random effects were retained.

Research question	Random effect	Variance	SD	ICC
1	Participant (intercept)	0.027	0.165	0.051
1	Participant (residual)	0.504	0.710	
5*	School (intercept)	<0.001	0.025	NA
6	Participant (intercept)	1.480	1.216	0.085
6	Participant (residual)	15.980	3.997	

Table A2. RQ 1: Does VBB reduce overall examiner variability?.

Term	Estimate	Std.error	t-value	p.value	Conf.low	Conf.high
(Intercept)	1.077	0.069	15.602	0.000	0.942	1.213
Group	−0.053	0.048	−1.096	0.273	−0.148	0.042
Performance	0.070	0.045	1.560	0.119	−0.018	0.158
School A	−0.045	0.162	−0.276	0.783	−0.362	0.272
School B	−0.175	0.168	−1.037	0.300	−0.505	0.155
School C	0.076	0.143	0.529	0.597	−0.205	0.356
School D	0.223	0.181	1.235	0.217	−0.131	0.577
School E	−0.076	0.172	−0.439	0.661	−0.413	0.262
School F	0.262	0.143	1.829	0.067	−0.019	0.542
School G	0.078	0.145	0.539	0.590	−0.206	0.363
School H	0.013	0.149	0.088	0.930	−0.279	0.305
School I	−0.155	0.406	−0.382	0.702	−0.950	0.640
School J	0.232	0.131	1.772	0.076	−0.025	0.488
School K	−0.259	0.176	−1.476	0.140	−0.603	0.085
Order	0.028	0.051	0.543	0.587	−0.072	0.128
Group x Performance	0.057	0.045	1.286	0.198	−0.030	0.145

Participant-level clustering was modest in the linear mixed models, with ICCs of 0.051 (Research question 1) and 0.085 (Research question 6), indicating that most variability in outcomes occurred within participants rather than between participants. For the model including School as a random effect (Research question 5), the estimated random intercept variance was very small (<0.001), indicating negligible differences in mean outcomes between schools. Residual variance in this model was modelled *via* a dispersion parameter, allowing it to vary across schools and allocations; therefore, a single ICC is not reported.

Model parameter estimates (i.e. main effects) for each model in analysis

Reference groups for analysis (applies to all tables:

- Group: Intervention
- Performance: Good
- School: L
- Order: 2

Table A3. RQ 2: Does VBB reduce examiner variability for outlying examiners?.

Term	Estimate	Std.error	t-value	p.value	Conf.low	Conf.high
(Intercept)	1.361	0.082	16.594	0.000	1.181	1.507
Group	0.000	0.060	0.004	0.997	−0.115	0.123
Performance	−0.068	0.058	−1.163	0.248	−0.190	0.047
School A	−0.195	0.248	−0.787	0.433	−0.865	0.214
School B	−0.552	0.301	−1.838	0.069	−1.460	−0.074
School C	−0.054	0.158	−0.345	0.731	−0.408	0.234
School D	−0.004	0.198	−0.021	0.983	−0.477	0.339
School E	−0.024	0.301	−0.080	0.936	−0.862	0.477
School F	0.134	0.127	1.057	0.293	−0.125	0.376
School G	−0.087	0.168	−0.516	0.607	−0.465	0.214
School H	0.246	0.177	1.395	0.166	−0.164	0.560
School I	0.316	0.143	2.206	0.030	0.021	0.583
School J	0.102	0.252	0.404	0.687	−0.570	0.518
Order	0.104	0.072	1.440	0.153	−0.029	0.249
Group x Performance	0.143	0.058	2.459	0.016	0.030	0.266

Table A4. RQ 3 does VBB increase the proportion of correctly classified performances?.

Term	Estimate	Std.error	z-value	p.value	Conf.low	Conf.high
(Intercept)	0.736	0.167	4.397	0.000	0.422	1.084
Group	−0.007	0.167	−0.042	0.966	−0.329	0.336
Performance	−1.679	0.169	−9.941	0.000	−2.032	−1.365
Order	0.241	0.147	1.645	0.100	−0.044	0.533
Group x Performance	−0.372	0.167	−2.222	0.026	−0.716	−0.052

Table A5. RQ4: Does VBB increase accuracy (i.e. distance from agreed 'correct' score)?

Term	Estimate	Std.error	t-value	p.value	Conf.low	Conf.high
(Intercept)	1.708	0.223	7.646	0.000	1.268	2.147
Group	−0.082	0.223	−0.366	0.714	−0.521	0.357
Performance	1.609	0.223	7.209	0.000	1.170	2.048
Order1	−0.036	0.223	−0.162	0.871	−0.476	0.403
Group x Performance	0.373	0.223	1.673	0.095	−0.066	0.812

Table A6. RQ5: Does VBB increase judgmental confidence?

Term	Estimate	Std.error	z-value	p.value	Conf.low	Conf.high
(Intercept)	1.652	0.020	80.991	0.000	1.612	1.692
Group	0.049	0.014	3.440	0.001	0.021	0.077

Table A7. RQ6: Are inter-school differences apparent in examiner scoring?

Term	Estimate	Std.error	t-value	df	p.value	Conf.low	Conf.high
(Intercept)	19.183	0.292	65.670	157	0.000	18.606	19.760
School A	0.567	0.806	0.703	157	0.483	−1.025	2.158
School B	0.516	0.814	0.635	157	0.527	−1.091	2.123
School C	0.492	0.709	0.693	157	0.489	−0.909	1.893
School D	0.252	0.900	0.280	157	0.780	−1.526	2.030
School E	−0.111	0.847	−0.130	157	0.896	−1.784	1.563
School F	−0.251	0.708	−0.355	157	0.723	−1.649	1.146
School G	−0.500	0.723	−0.691	157	0.491	−1.929	0.929
School H	−0.221	0.742	−0.298	157	0.766	−1.686	1.244
School I	0.257	2.031	0.126	157	0.900	−3.755	4.268
School J	0.363	0.646	0.563	157	0.575	−0.912	1.638
School K	−0.633	0.873	−0.726	157	0.469	−2.357	1.091
Group	−0.097	0.239	−0.407	157	0.685	−0.568	0.374
Performance	−0.891	0.216	−4.122	169	0.000	−1.318	−0.464
Order	−0.037	0.249	−0.150	157	0.881	−0.528	0.454
Group x Performance	0.373	0.216	1.727	169	0.086	−0.054	0.800