Article:
Homer, M. and Ababei, V. (Accepted: 2026) Evidencing improvement in examiner calibration in OSCEs. Medical Teacher. ISSN: 0142-159X (In Press)

Evidencing improvement in examiner calibration in OSCEs

Authors

Matt Homer, Schools of Education and Medicine, University of Leeds, LS2 9JT, UK
m.s.homer@leeds.ac.uk

Vlad Ababei, Assessment Development Team, General Medical Council, 3 Hardman Street, Manchester, M3 3AW, UK

Abstract

In developing and administering OSCE-type assessments, institutions can spend significant resources training examiners and designing stations/scoring instruments in attempts to ensure that they are well-calibrated – for example, across parallel circuits where the same station is administered across nested groups of candidates. This paper, situated in the context of a high-stakes OSCE for international medical graduates wanting to work in the national health service in the UK (data over the period 2016-2024), employs a recently developed quantitative measure of examiner calibration to identify which stations show relatively high and low average degrees of calibration between examiners. Using documentary analysis of station material (e.g. marksheets and other supporting information), we then investigate these stations qualitatively to better understand what factors might drive better calibration as stations develop and design elements changes over time. We find that those stations that are better calibrated are typically newer with more detailed and relevant scoring guidance and support materials, whilst there is little evidence that the nature of the task(s) or other contextual factors (e.g. simulated patient age, sex or ethnicity) are important in determining calibration levels. In this work we provide strong evidence of how key developments in station design, the quality of support materials and enhanced examiner training practices can succeed in improving degrees of calibration in OSCE stations – and suggest ways that all institutions might improve their practices in this regard.

Key words

OSCEs; calibration; parallel circuits; support materials

Practices points

- Minimising differences in patterns of OSCE examiner scoring, for example across parallel circuits, is important in assuring the quality of assessment outcomes
- There are a range of efforts made to improve calibration between examiners, including specific training, the development of a range of appropriate written support materials at the station-level, and other calibration practices in OSCE stations
- Using a recently developed quantitative measure of examiner alignment, this study shows how improved calibration practices over time can enhance degrees of calibration in stations
- This work adds to the relatively sparse literature on 'what works' to minimise differences in examiner scoring patterns in OSCEs

Highlights

This paper evidences how developments in calibration practices and support materials in OSCEs can lead to improved alignment in scoring between OSCE examiners thereby increasing confidence in assessment outcomes.

Biographical note

Matt Homer is an academic who works in both the Schools of Education and Medicine at the University of Leeds. He has a long-standing interest in improving the quality of medical education assessment, usually via quantitative and psychometric investigations. He has published widely in areas such as standard setting and examiner stringency in OSCEs, and has external assessment advisory roles with a range of institutions including the General Medical Council in the UK.

Vlad Ababei is an Assessment Officer for the General Medical Council in the UK, working within the Clinical Assessment Centre (CAC) in Manchester. He is responsible for developing, curating and continuously improving content for the exam for international medical graduates who want to work as doctors in the UK National Health Service (commonly known as PLAB), as well as other GMC-regulated assessments. He has an interest in people management and academic research in assessments.

Ethics, funding and data access

On registering for the exam, PLAB candidates consent to their examination data being used for research purposes and to allow for improvements in exam practice. The General Medical Council in the UK gave permission for the analysis of the anonymized data used in this paper. This research was not externally funded but the lead author acts as a paid assessment/psychometric consultant to the PLAB2 panel on an *ad hoc* basis. For reasons of confidentiality and assessment security, the raw data used in this study is not available to other researchers.

Declaration of interest

We have no declarations of interest to report.

Introduction

What to do about examiners?

There is a lot of evidence in medical education and other assessment literature that ratings of performance can vary by assessors in important ways. Investigations and theories around the possible sources of such variation include, for example differing frames of reference that examiners might have (Yeates et al., 2013), rater/examiner drift (Harik et al., 2009) and time/contrast effects (Hope and Cameron, 2015; Yeates et al., 2022). Evidence also suggests that judging levels of performance is cognitively difficult for examiners, particularly at lower levels of performance (Malau-Aduli et al., 2021) and that levels of examiner stringency can vary in individuals to an extent depending on the nature of the scoring instrument (Homer, 2024).

When the design of an OSCE requires parallel circuits, any (unwanted) examiner variation across circuits is usually challenging to estimate, but can impact in different ways on different groups of candidates leading to unfair outcomes (Swanson et al., 2013; Yeates et al., 2021). Hence, in high stakes summative settings, it is important for examiners to be as well-aligned in their scoring as possible with station design, support materials and appropriate examiner training intended to facilitate good calibration across circuits (Khan et al., 2013; Harden et al., 2015, ch. 9; General Medical Council, 2024). Specific station development guidance in the literature suggests that well-constructed marking sheets, focussed training in these can help improve levels of calibration between examiners (Moreno-López and Sinclair, 2020; Malau-Aduli et al., 2023) as can video-based benchmarking for examiners (Edwards et al., 2025). However, there remain important questions around the impact of feedback to examiners on the quality and consistency of their marking (Sturman et al., 2017; Crossley et al., 2019).

In summary, there is a lack of clear evidence, particularly longitudinal, as to what might 'work' to meaningfully improve degrees of examiner alignment in performance assessments such as OSCEs – despite an extant range of best practice guidance and understandings of potential sources of variation in scoring.

This paper – what makes calibration between examiners better?

In this study, a recently developed measure of variation (i.e. level of consistency) between examiner scores (Homer, 2025)[1] is used to identify specific stations that have relatively low or high levels of consistency on this metric across many station administrations. Using qualitative methods, including documentary analysis (Bowen, 2009), we investigate what features these stations have in common and how they differ. Our aim is to develop insight into the nature of station characteristics, associated support materials and training that impacts on successful (or otherwise) calibration in scoring between examiners at the station level.

Our data derives from PLAB2, a summative OSCE testing clinical and professional skills, knowledge and behaviours for those international medical graduates who want to come to work in the National Health Service in the UK. Full contextual details of the exam are available elsewhere (Homer, 2024; General Medical Council, 2025) but, in essence, this is a 16-station OSCE intended to reflect real life settings such as consultation with a general practitioner or day-to-day clinical activities appropriate for a foundation doctor a year on from completing an undergraduate medical degree. In its current 2025 format, candidates are assessed by a clinically trained examiner, and most stations are administered across two parallel circuits

[1] This new measure will be described in detail in the Methodology section.

with examiners seeing two groups of candidates (i.e. ≈32 candidates in total each). PLAB2 exams take place regularly at two separate sites in Manchester, UK throughout the year – for example, in 2024 there were 344 separate PLAB2 exams including almost 5500 individual station administrations.

In each station, candidate performance is scored with a global grade (0=*fail*, 1=*borderline*, 2=*satisfactory*, 3=*good* and a total domain score (0 to 12) across three domains (*Data gathering, technical and assessment skills*; *Clinical management skills*; and *Interpersonal skills* – all scored 0 to 4) (General Medical Council, 2025). Borderline regression (McKinley and Norcini, 2014) is used for standard setting in PLAB2.

The paper continues with details of the sequential mixed methods methodology (Ivankova et al., 2006) we employ in the study, and then the quantitative and qualitative findings are presented in turn. We take a pragmatic approach to our research and choice of appropriate methods (Foster, 2024) – based on consideration of what types of evidence we can best use to achieve our research aims. We finish the paper with discussion of what this work adds to what is already known about how best to calibrate examiners/stations and to improve OSCE scoring practices in general.

Methodology

We detail first the quantitative phase of the study – outlining how we identify those stations with the highest and lowest levels of misalignment. This is followed by a description of the second phase - a documentary analysis of station support materials intended to reveal commonalities and differences across these stations.

Quantitative phase: identifying stations with the highest and lowest consistency in examiner scoring

During the period November 2016 to October 2024 there were 10,226 separate station administrations of PLAB2 with exactly two examiners in parallel circuits. This is just under half of all station administrations over this period, with the majority of the remaining administrations with only a single examiner. PLAB2 has a large station bank and there are 825 different stations represented in the two-circuit dataset, with a minimum station occurrence 1, maximum 70 and median 8.

To give us confidence in the robustness of our station-level analysis, we limit our investigation to the subset of stations present at least 10 times in the two-circuit data. This subset consists of 8,238 station administrations in total, across 382 individual stations with median occurrence of a particular station=18.

For each station administration, the examiner-level borderline regression slope and intercept (Pell et al., 2010; McKinley and Norcini, 2014) are the key metrics we use to calculate levels of misalignment across the two circuits. To do this, we calculate the area between the two borderline regression lines for pairs of examiners in each separate station administration (Homer, 2025). An exemplar of the hypothetical situation is shown in Figure 1 with a blue and orange circuit and the grey shaded area showing the level of misalignment between examiners for this hypothesised station. Each point in the figure corresponds to a single candidate.

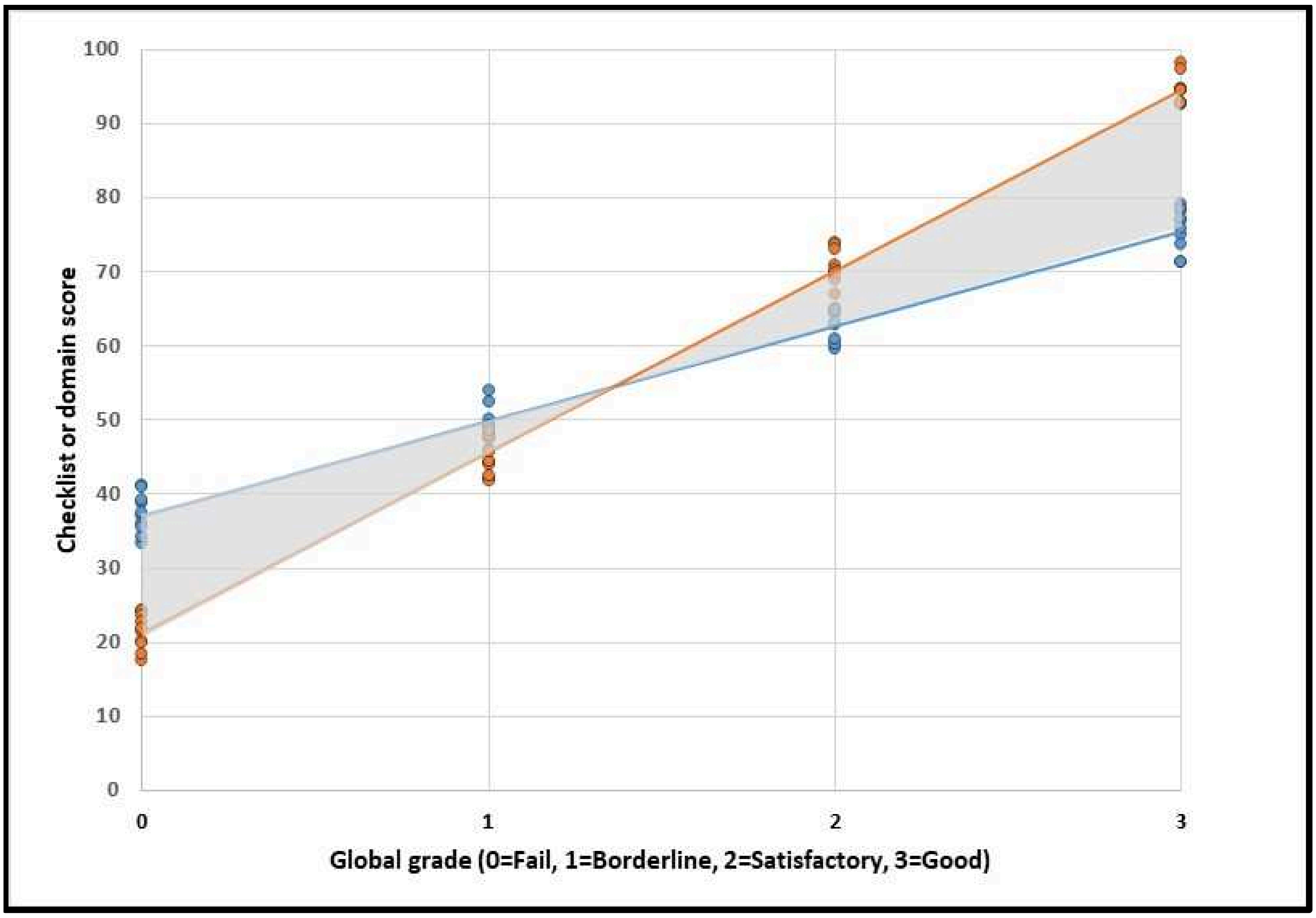

Figure 1: Scatter plot of grades versus domain scores within in station (area between examiner regression lines shaded)

The area value, expressed as a percentage of the total area in the scatter plot, can be conceptualised as measuring the degree of misalignment between examiners – with a 'large' area corresponding to relatively poor calibration, and a small area (i.e. near 0) indicating good calibration where the pair of regression lines are almost completely overlapping. The key innovation in this metric is that it employs both scores that examiners award at the station level *simultaneously* (i.e. the global grade and the total domain score) (Homer, 2025).

We then use a simple variance component analysis (Bloch and Norman, 2012), treating station as a random effect, to estimate the proportion of variation in the area measure explained by the station factor. This gives us an indication across the full dataset of how much variation by individual station there is in this measure.

We also produce the average (median or mean) area for each station across our data and can then identify stations that are typically well-calibrated (i.e. low average area) or those less well-calibrated (i.e. high average area). We present the top and bottom 10 such stations (see Table 1 in the Findings), which we then investigate further in the second, qualitative, part of the study.

Qualitative phase: documentary analysis of station materials

For those twenty stations identified in the quantitative phase, we analyse the corresponding range of PLAB2 support materials and practices, mostly at the station level – such as documents containing scripts for simulated patients (SPs) and examiners, and grade descriptors within stations. These materials are used by examiners and SPs during pre-exam calibration discussions that take place on the morning of each PLAB2 exam between those in the same station, but different circuits. These discussions are intended to ensure that examiners and SPs across parallel circuits have a shared understanding of expectations around what activities are important to the station, how different levels of performance of candidates should be scored, and key elements of the SP performance.

Our main methodological approach in this phase is that of documentary analysis in three parts - skimming (superficial examination), reading (thorough examination), and interpretation (Bowen, 2009). This analysis, carried out by VA with full knowledge of data in Table 1 , was also informed by our knowledge of broader contextual PLAB2 factors that have developed over time since 2016. These include practices related to examiner training and support such as the exact nature of the calibration procedures on the day of the exam, the post-exam feedback system to examiners and changes to the annual examiner appraisal systems. We will return to some of these elements in the later parts of the paper.

We began the documentary analysis by skimming over the documents related to each of the twenty stations to gain a general sense of content, purpose and structure of each. In hard copy, these documents are typically made up of:

- a page of candidate instructions
- three to five pages of SP instructions
- two to four pages of marking criteria
- a page of examiner calibration guidance
- one or two pages of supporting information (for example, material from medical reference websites)
- one to three pages in the form of laminates for stations where there are clinical test results to be handed over to the candidate at the appropriate time

In total, this material amounts to between nine and 15 pages per station.

The documentary analysis led to the initial development of overarching common factors across the documents, and to classifying the quality and status of station-level content into categories. These factors and categories were further developed through more detailed reading of the documents. In the final interpretation stage, we focussed on identifying the key differences observed in this process between the relatively well-calibrated stations and those with lower levels of scoring alignment between examiners.

During this analysis, we found five main factors revealing important differences across the two groups of stations. Two of these, both related to station materials/content, leant themselves to additional qualitative analysis so we also devised a rating scale (1=adequate, 2=reasonable or 3=good) for each sub-factor (Morgan, 2022), in order to better interrogate factors such as: quality of writing, clarity, accuracy, consistency and level of detail of the information provided, as well as alignment to current station writing guidance and best practices (see Table 2 and Table 3 in the Appendix for more details).

.

Findings

We begin with the quantitative analysis of misalignment/calibration and then move on to the documentary analysis of station support and other materials.

Quantitative results: variation in average misalignment by station

Figure 2 shows how the area (percentage) metric varies by individual station administration (n=8,238, median=5.68, mean=6.52, 5th percentile=1.40,95th percentile=14.66). Those station administrations on the left of the histogram are well-calibrated with low levels of misalignment across circuits (i.e. the area between the two regression lines is small – see Figure 1). The opposite is true of those on the right of the distribution.

Typically (i.e. across the full data), the area between regression lines is about 6% of the area of the full scatter plot of global grades (x) versus checklist/domain scores (y).

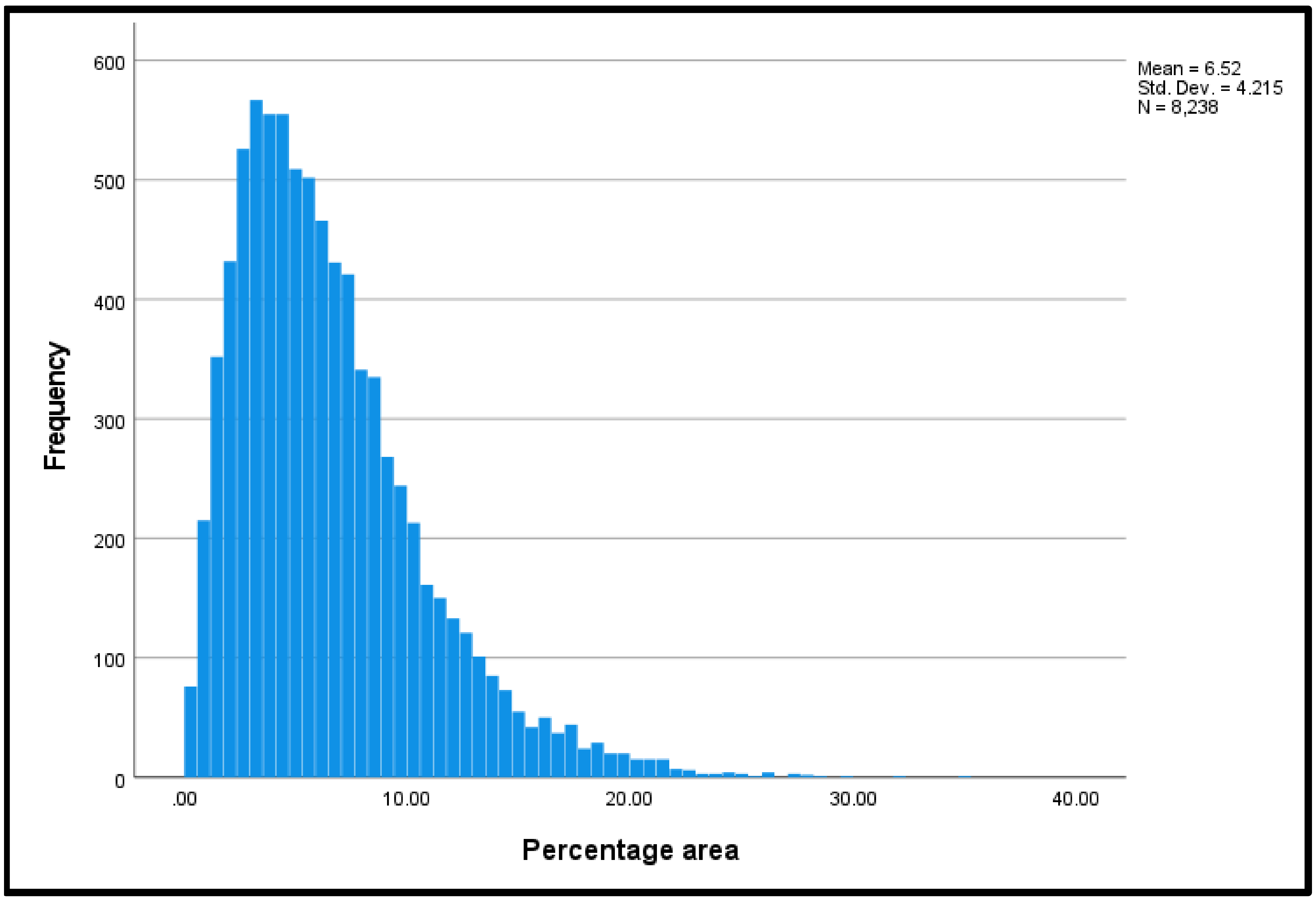

Figure 2: Histogram of percentage area metric across station administrations

A variance component analysis suggests that only 1.5% of the variation in area is due to individual stations. However, when aggregating across multiple administrations of the same station we do find some important differences in degrees of calibration by station. The stations listed at the top of Table 1 are those that are the least well-calibrated across multiple administrations on our measure (i.e. with higher average area), and those at the bottom are those that are the best calibrated.

We see that the degree of misalignment of those stations towards the top of Table 1 is more than twice those towards the bottom.

Station pseudo-identifier	Station title	Percentage area		Number of station administrations
		Mean	Median	
1	Patient complaint	11.75	9.67	10
2	Heart attack	10.56	11.89	11
3	Knee examination	9.78	9.33	16
4	Teaching a student doctor	9.75	9.06	11
5	Man attending follow up appointment	9.42	9.58	18

6	Vaccination	9.42	7.81	10
7	Teaching a medical student	9.36	7.33	10
8	Parent with concerns about their child	9.22	6.92	31
9	Relatives' requests	9.17	6.81	20
10	Deafness	8.92	8.36	10
373	Clostridium difficile	4.44	3.06	10
374	Swelling	4.44	3.72	15
375	Boy with sore throat	4.42	4.39	10
376	Osteoporosis	4.36	4.28	12
377	Worried parent	4.28	3.11	10
378	Stomach pain	4.17	4.81	14
379	49 year old patient attending appointment	4.14	3.25	22
380	Emergency GP appointment	3.94	2.64	10
381	GP appointment	3.94	3.19	12
382	Appointment to discuss concerns	3.75	2.67	11

Table 1: The highest and lowest 10 stations on mean/median area metric

The actual station identifier (not shown in Table 1) can be used as a proxy for the period of development of the station. A correlation between this identifier and area metric shows a negative relationship (r=-0.089, p<0.001, n=8,238) which, in a simple analysis, suggests that calibration levels in stations have improved over time when comparing older stations to those more recently developed.

In part, this finding of apparent improvements in calibration over time motivates the documentary analysis of station materials in Table 1 that follows.

Qualitative results: Identifying important calibration factors in stations

We now discuss the five factors that emerged in the documentary analysis. To maintain exam security, findings are referred to in general rather than discussing specific detailed elements of the station content. A summary of our documentary analysis for each of the stations is shown in the Appendix (see Table 2 for summary of the actual documentary analysis results, and Table 3 for the scoring guidance used during this process).

1. Marking criteria in the station

Arguably the biggest factor identified in the documentary analysis is the marking criteria element of the station support materials where a range of improvements to PLAB2 examiner scoring guidance have been made over the period 2016 to 2024. Originally, for each station and each of the three domains (*Data gathering, technical and assessment skills*; *Clinical management skills; and Interpersonal skills),* there were positive and negative descriptors of performance. These covered the extremes of a candidate's performance, i.e. what a very good and a very bad candidate would do. In the newer 'ACE descriptor' marksheets, these old levels align with the 'A' and 'E' candidate descriptors. In addition, this new marking structure introduced the 'C' candidate, a set of marking criteria defining an adequate candidate who is deemed safe in their approach and patient management (so is set just above 'borderline').

ACE descriptors were developed to bolster the original marking criteria, and stations were slowly converted over time to the new, improved scheme.

Whilst most of the 20 stations in Table 1 now have ACE descriptors, the group at the top typically only had ACE descriptors added more recently. Eight of this group had positive/negative descriptors marksheets at some point in time, whereas the corresponding figure for the well-calibrated group is only three out of the ten. This analysis suggests that adding the full set of ACE descriptors to the stations has contributed to improved calibration statistics by offering examiners more information to help them discriminate between excellent (A), adequate (C) and very poorly performing (E) candidates.

2. *Station content and other factors*

Development work has been carried out over time to improve a range of station-level materials including the SP script, examiner calibration guidance and examiner (medical) supporting information. We take each in turn.

Simulated patient script

These materials are intended to provide the SP with the necessary history and guidance on the patient they are portraying. The way these have been written has developed over time in PLAB2. For example, the order of the information presented has been standardised, set phrases have been introduced to make scripts more consistent across stations, and how SPs are to react to unforeseen questions has been developed.

A review of these materials suggests that in the well-calibrated stations:

- The information is valid, relevant, well-organised and follows the agreed order currently deemed best practice by the PLAB2 assessment team (i.e. in order: presenting issue; past medical history; medication; allergies; family medical history; social history - diet, exercise, alcohol, smoking, living conditions etc; any other relevant information).
- Adequate details of the patient's 'story' are present at a level sufficient in helping facilitate the candidate-patient encounter, and to ensure that SP performance is as consistent as possible across paired circuits.
- Phrasing and grammar is accurate and clear.

By contrast, the less well-calibrated stations do not follow the best-practice guidelines and PLAB's 'style guide'[2] in at least some of these areas.

Examiner calibration guidance

These support materials give specific suggestions for the examiners regarding what they should focus on during the calibration discussion that takes place immediately prior to the exam – involving both examiners and SPs in the same station, but different circuits. For example, in some stations this guidance will cover how to react to a candidate's request to examine the patient – and will align in this regard with the SP script. This guidance has developed over time, for example, it now states that they must always complete a full run

[2] These are documents that guide station writers to maximise standardisation across station materials. They formalise what kinds of words to use, how numerical values are presented, give general formatting rules, and record particular policy decisions around how things should be presented in stations.

through of the station with one of them acting as a candidate as part of the calibration session.

The calibration guidance material is not intended as an exhaustive list or a ‘box-ticking’ exercise and there is an expectation that examiners will use their professional experience to guide the calibration process. However, in some of the less well-calibrated stations at the top of Table 1, our documentary analysis shows the calibration guidance was either limited, missing altogether or perhaps too generic (i.e. without station-specific details).

Examiner (medical) supporting information

This material is summarised medical information designed to expand on the context and conditions relevant to the station. In well-calibrated stations, our analysis suggests that the information:

- Is from doctor-focused sources that provide clear and authoritative clinical guidance - for example, the GMC’s Good Medical Practice[3] and National Institute for Health and Care Excellence guidelines[4] in the UK) - rather than from patient-oriented sources (for example, public-facing NHS websites).
- Is of appropriate length and always provides context on all relevant aspects of the station (for example, covering both history taking and management elements when appropriate).
- Is sufficiently specific and clear in terms of exactly how the station should run.

By contrast, the less well-calibrated stations tend to have elements of content and medical support materials that are a little disorganised, not always sufficiently clear or are lacking in detail, and sometimes include irrelevant information.

3. *Year of the initial station development*

Typically, the stations in the less well-calibrated group (top of Table 1) are older (median year of writing is 2017 compared to 2018 in the well-calibrated). Whilst the year itself does not necessarily mark any specific major change in the PLAB2 operation or policy, a range of PLAB2 training and quality assurance processes have been implemented with the intention of improving station quality over time. Our analysis suggests that this focus on continuous improvement has had a positive impact over time in calibration practices. All stations are under ongoing scrutiny, with regular quality assurance checks, audits and formal feedback from examiners and SPs on station performance. Examiners and SPs are encouraged to raise concerns and suggest improvements for the stations they are assigned to on a given exam day – for example, reporting instances where a common question every candidate asks on a specific day is not covered in the scripts.

This finding algins with the (negative) correlation identified in the quantitative analysis across all stations between station identifier (a proxy for year of development) and the measure of calibration.

4. *Station status (live or archived)*

[3] https://www.gmc-uk.org/professional-standards/the-professional-standards/good-medical-practice
[4] https://www.nice.org.uk/what-nice-does/our-guidance/about-nice-guidelines

Stations deemed to be performing poorly in post-exam psychometric review or those which are no longer medically applicable are taken out of use in the live exam station bank as part of the continuous quality assurance processes. In Table 1, all the well-calibrated stations at the bottom are 'live' but only five of the 10 less well-calibrated stations are currently 'live'. We can conclude that prior to the availability of the new calibration metric, quality control procedures have tended to identify and remove less well-calibrated stations from the PLAB2 'live' station bank.

5. *Current station version (1 or>1)*

Station materials are sometimes modified based on feedback from a number of stakeholders including examiners, SP facilitators[5] and the delivery/administrative team in order to improve them for future administrations. Every time a minor amendment to station materials is made, but where the fundamental activity being assessed is deemed unchanged, the station version number is increased by one[6]. This usually occurs when feedback received from an examiner, SP or observer suggests a modification will likely improve station performance.

Eight out of 10 of well-calibrated stations in Table 1 have been amended at least once whereas only 3 out of 10 of those in the poorer calibrated group have been amended (see Table 2). This suggests that changes made to stations - based on feedback or active quality assurance checks – have tended to improve examiner calibration in stations.

Discussion

This paper set out to investigate 'what works' in terms of improving calibration between examiners in OSCEs via improved design, materials and related practices - given all the potential sources of differences in examiner scoring that are known to exist in OSCEs (Harik et al., 2009; Yeates et al., 2013; Hope and Cameron, 2015; Homer, 2024) and what is known about how various elements of OSCE design and practice can help minimise such differences (Khan et al., 2013; Harden et al., 2015; Moreno-López and Sinclair, 2020; Malau-Aduli et al., 2021; General Medical Council, 2024).

Using new quantitative methods to identify how well stations have been calibrated (Homer, 2025), we systematically analysed station materials to develop understanding of what particular changes to PLAB2 exam practices seem to result in better calibration metrics. At the 'big-picture' level, this review of materials highlights an obvious difference in overall quality between the two sets of stations (Table 1) in terms of all the station-level support materials and wider PLAB2 quality assurance practices in place. For newly developed stations, scripts are better organised and the level of detail within them has been improved over time to better support SPs and examiners alike. In older stations, this is not always the case. Across all materials, the standardisation of language used as well as the phrasing and structure have also improved – and are now carefully prescribed in station development rules and guidance, which themselves remain subject to continuous ongoing improvement. This work adds, therefore, to the relatively limited evidence base that details specific tools and enhancements that can aid examiner scoring consistency and alignment in OSCE-type assessments (Khan et al., 2013; Harden et al., 2015, ch. 9; Malau-Aduli et al., 2023).

[5] These are senior SPs who observe stations throughout the day via a video feed and ensure SPs perform as expected.

[6] In cases where substantive changes are made to a station, for example, in producing a cloned version where key patient demographics have changed, the station is given a new and unique identifier).

A key element found to improve scoring alignment in this work is in the development of the ACE marking descriptors which offer examiners a more detailed and consistent framework for evaluating a range of candidate performance. Arguably, this is an unsurprising result as it is consistent with practitioner guidance in this regard (Streiner and Norman, 2008, ch. 7; Yudkowsky, 2019, ch. 7) and aligns with work suggesting that 'borderline' levels of performance can be challenging to judge accurately (Malau-Aduli et al., 2021) . As a consequence of this new evidence, the PLAB2 development team has committed to updating approximately 500 existing OSCE station marksheets to align with the ACE descriptors model – as of late 2025 this work is largely complete.We know that OSCE-type assessments are expensive to administer (Pell et al., 2013; Harden et al., 2015, ch. 15), and this research does underline the importance of adequately resourcing the ongoing development of any high-stakes OSCE. The improvement of OSCE materials and guidance to a consistently high standard requires suitably qualified staff to drive ongoing developments, and to lead the training and development of key stakeholders including station writers and OSCE examiners – particularly when it comes to best practices around calibration and marking guidance (Khan et al., 2013; General Medical Council, 2024).

The focus in the literature, and to an extent also in this work, tends to assume that the main source of calibration issues is due solely to examiner behaviours. However, one important finding in our work is the emergence of the quality of SP scripts as really helping in improving calibration. In short, part of the source of 'examiner' variation can be due to differences in how SPs play the role across parallel circuits, and improved guidance on this (e.g. making it valid, relevant, accurate, appropriately storied) can help to improve calibration. The role of the SP in influencing differences OSCE in scoring certainly deserves greater scrutiny.

Another important emergent in our findings is the medical information provided to examiners – we know that clinical expertise and experience can impact on judgments of performance (Yeates et al., 2013). Our work implies that making explicit what is regarded as 'correct' in the scenario – both in terms of medical knowledge and interpretation of behaviours - is likely to improve scoring alignment – presumably by helping to weaken any influence between clinical experiences examiners might bring and their scoring of performance.

Our documentary analysis work also suggests that some station elements appear to have little to no impact on the degree of calibration observed – at least in our context. The type of scenario in the station, the theme addressed within it, and patient characteristics did not emerge as important influences on the degree of marking alignment between examiners on parallel circuits during the analysis – but there is the need for more systematic and wider-scale research in this area.

In terms of recommendations for practice derived from our study, we would suggest that institutions pay close attention to ongoing development of all the OSCE support materials, in particular the marking criteria in the station at the middle and lower end of the performance spectrum, the SP script, and the specific up-to-date medical knowledge required in each station. With parallel circuits, also providing clear guidance to examiners around key elements of the station to discuss during pre-exam calibration discussion is likely to be effective.

Study limitations and final thoughts

The obvious limitation of this study is that it is from a single specific exam context, with necessarily limited sampling of 20 stations for the documentary analysis – which was carried out by a single analyst (VA). We would also note that because our study is observational in nature, formally we cannot make strong causal claims. Ideally, experimental methods would be needed for this, but these are challenging to carry out – particularly longitudinally. However, our work does add to the seemingly sparse published research into the wider area of longitudinal OSCE quality improvement (Fuller et al., 2013; Boursicot et al., 2021). There is also the sense that this work validates the application of the new calibration metric via a mixed methods study (Homer, 2025; Supianto, 2025) – in essence we have measured something (degrees of calibration) made efforts to improve things (station quality in its broadest sense) and the measure does then improve.

We conclude by emphasising the difficulty of researching and publishing assessment-related research at the appropriate level of detail useful to readers, whilst also maintaining sufficient levels of exam security around specific station content – and hope in this work we have achieved the right balance to make it useful to readers and practitioners in a range of OSCE contexts.

References

Bloch, R. and Norman, G. 2012. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher*. **34**(11), pp.960–992.

Boursicot, K., Kemp, S., Wilkinson, T., Findyartini, A., Canning, C., Cilliers, F. and Fuller, R. 2021. Performance assessment: Consensus statement and recommendations from the 2020 Ottawa Conference. *Medical Teacher*. **43**(1), pp.58–67.

Bowen, G.A. 2009. Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*. **9**(2), pp.27–40.

Crossley, J.G.M., Groves, J., Croke, D. and Brennan, P.A. 2019. Examiner training: A study of examiners making sense of norm-referenced feedback. *Medical Teacher*. **41**(7), pp.787–794.

Edwards, R.J., Yeates, P., Lefroy, J. and McKinley, R. 2025. Understanding contexts and mechanisms through which video based benchmarking promotes alignment of examiners' scoring in objective structured clinical exams. *Advances in Health Sciences Education*.

Foster, C. 2024. Methodological pragmatism in educational research: from qualitative-quantitative to exploratory-confirmatory distinctions. *International Journal of Research&Method in Education*. **47**(1), pp.4–19.

Fuller, R., Homer, M. and Pell, G. 2013. Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. *Medical Teacher*. **35**(6), pp.515–517.

General Medical Council 2024. *Requirements for the MLA Clinical and Professional Skills Assessment* [Online]. London: GMC. [Accessed 10 September 2025]. Available from: https://www.gmc-uk.org/-/media/documents/mla-cpsa-requirements-_pdf-84742729.pdf.

General Medical Council 2025. What is the PLAB 2 exam? *What is the PLAB 2 exam?*. [Online]. [Accessed 7 May 2020]. Available from: https://www.gmc-uk.org/registration-and-licensing/join-our-registers/plab/plab-2-guide/what-is-the-plab-2-exam.

Harden, R., Lilley, P. and Patricio, M. 2015. *The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment., 1e* 1 edition. Edinburgh ; New York: Churchill Livingstone.

Harik, P., Clauser, B.E., Grabovsky, I., Nungester, R.J., Swanson, D. and Nandakumar, R. 2009. An Examination of Rater Drift Within a Generalizability Theory Framework. *JOURNAL OF EDUCATIONAL MEASUREMENT*. **46**(1), pp.43–58.

Homer, M. 2025. Going beyond hawks and doves – Measuring degrees of examiner misalignment in OSCEs. *Medical Teacher*. **0**(0), pp.1–7.

Homer, M. 2024. Towards a more nuanced conceptualisation of differential examiner stringency in OSCEs. *Advances in Health Sciences Education*. **29**(3), pp.919–934.

Hope, D. and Cameron, H. 2015. Examiners are most lenient at the start of a two-day OSCE. *Medical Teacher*. **37**(1), pp.81–85.

Ivankova, N.V., Creswell, J.W. and Stick, S.L. 2006. Using Mixed-Methods Sequential Explanatory Design: From Theory to Practice. *Field Methods*. **18**(1), pp.3–20.

Khan, K.Z., Gaunt, K., Ramachandran, S. and Pushkar, P. 2013. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: organisation & administration. *Medical Teacher*. **35**(9), pp.e1447-1463.

Malau-Aduli, B.S., Hays, R.B., D'Souza, K., Saad, S.L., Rienits, H., Celenza, A. and Murphy, R. 2023. Twelve tips for improving the quality of assessor judgements in senior medical student clinical assessments. *Medical Teacher.*, pp.1–5.

Malau-Aduli, B.S., Hays, R.B., D'Souza, K., Smith, A.M., Jones, K., Turner, R., Shires, L., Smith, J., Saad, S., Richmond, C., Celenza, A. and Gupta, T.S. 2021. Examiners' decision-making processes in observation-based clinical examinations. *Medical Education*. **55**(3), pp.344–353.

McKinley, D.W. and Norcini, J.J. 2014. How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher*. **36**(2), pp.97–110.

Moreno-López, R. and Sinclair, S. 2020. Evaluation of a new e-learning resource for calibrating OSCE examiners on the use of rating scales. *European Journal of Dental Education*. **24**(2), pp.276–281.

Morgan, H. 2022. Conducting a Qualitative Document Analysis. *The Qualitative Report*. **27**(1), pp.64–77.

Pell, G., Fuller, R., Homer, M. and Roberts, T. 2013. Advancing the objective structured clinical examination: sequential testing in theory and practice. *Medical Education*. **47**(6), pp.569–577.

Pell, G., Fuller, R., Homer, M. and Roberts, T. 2010. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. *Medical Teacher*. **32**(10), pp.802–811.

Streiner, D.L. and Norman, G.R. 2008. *Health measurement scales: a practical guide to their development and use* 4th edn. Oxford; New York: Oxford University Press.

Sturman, N., Ostini, R., Wong, W.Y., Zhang, J. and David, M. 2017. 'On the same page'? The effect of GP examiner feedback on differences in rating severity in clinical assessments: a pre/post intervention study. *BMC MEDICAL EDUCATION*. **17**, p.101.

Supianto 2025. Refining examiner alignment measures in OSCEs: A call for broader application and deeper insight. *Medical Teacher*. **47**(9), pp.1558–1559.

Swanson, D.B., Johnson, K., Oliveira, D., Hayes, K. and Boursicot, K.A. 2013. Estimating the Reproducibility of OSCE Scores When Exams Involve Multiple Circuits.

Yeates, P., Moult, A., Cope, N., McCray, G., Fuller, R. and McKinley, R. 2022. Determining influence, interaction and causality of contrast and sequence effects in objective structured clinical exams. *Medical Education*. **56**(3), pp.292–302.

Yeates, P., Moult, A., Cope, N., McCray, G., Xilas, E., Lovelock, T., Vaughan, N., Daw, D., Fuller, R. and McKinley, R.K. (Bob) 2021. Measuring the Effect of Examiner Variability in a Multiple-Circuit Objective Structured Clinical Examination (OSCE). *Academic Medicine*. **96**(8), p.1189.

Yeates, P., O'Neill, P., Mann, K. and Eva, K. 2013. Seeing the same thing differently. *Advances in Health Sciences Education*. **18**(3), pp.325–341.

Yudkowsky, R. (ed.). 2019. *Assessment in Health Professions Education* 2nd edition. New York: Routledge.

Appendix

Table 2 shows a summary of the findings of the documentary analysis - the rating scale is intended to score materials from adequate (1) to good (3).

Table 3 gives the specific scoring guidance for each station factor.

					Examiner's marking criteria			Station content and other factors		
Station group	Station pseudo-identifier	Station status	Year of initial station development	Current station version	Marksheet rating	ACE complete?	ACE rating	SP Script rating	Examiner's Calibration Guidance rating	Examiner (medical) supporting information rating
High mean area = poorer calibration	1	Live	2018	1	Replaced[i]	Yes	2	1	1	1
	2	Live	2016	2	Replaced	Yes	2	2	1	2
	3	Removed	2016	1	Replaced	N/A[ii]	N/A	N/A	N/A	N/A
	4	Rested[iii]	2018	1	Replaced	Yes	2	2	3	2
	5	Live	2018	2	2	No	N/A	1	1	1
	6	Rested	2017	1	Replaced	Yes	2	3	2	3
	7	Removed	2018	1	Replaced	N/A	N/A	N/A	N/A	N/A
	8	Live	2017	1	1	No	N/A	2	1	1
	9	Removed	2016	1	Replaced	N/A	N/A	N/A	N/A	N/A
	10	Live	2016	5	Replaced	Yes	3	3	3	3
Low mean area =	373	Live	2016	1	N/A	Yes	2	1	2	3

Station group	Station pseudo-identifier	Station status	Year of initial station development	Current station version	Examiner's marking criteria			Station content and other factors		
					Marksheet rating	ACE complete?	ACE rating	SP Script rating	Examiner's Calibration Guidance rating	Examiner (medical) supporting information rating
better calibration	374	Live	2018	2	2	Yes	2	3	3	2
	375	Live	2018	2	Replaced	Yes	3	3	3	3
	376	Live	2016	2	N/A	Yes	3	3	2	3
	377	Live	2017	4	3	No	N/A	3	3	3
	378	Live	2022	1	N/A	Yes	3	3	3	3
	379	Live	2022	5	N/A	Yes	3	3	3	3
	380	Live	2018	4	Replaced	Yes	3	3	3	3
	381	Live	2019	2	N/A	Yes	2	2	3	2
	382	Live	2020	4	Replaced	Yes	3	3	3	3

i: 'Replaced' implies the station was originally developed having positive-negative marksheets which were eventually replaced with ACE marksheets

ii: Not Applicable / does not exist

iii: 'Rested' means the station is temporarily removed from live exam for review/editing

Table 2: A summary of the documentary analysis for the 20 stations from Table 1

Marksheets	
1	Small number of marking criteria for each domain (2-4); broad/vague individual criteria, which do not make explicit what the candidate is expected to do (e.g. 'Takes history').
2	Small to average number of individual criteria (3-5); some of the marking criteria have examples of what is expected from the candidate (e.g. 'Takes history of the pain - onset, location etc.').
3	Good amount of marking criteria (4+) with appropriate level of detail; examples are used for most marking criteria, and it is clearer what the candidate is expected to do; covers extremes of performance well.
ACE	
1	Small number of marking criteria (2-3) per candidate level (A, C&E); marking criteria is vague/broad and lacks examples; discrimination between A and C candidates is minimally described.
2	Average number of marking criteria (3-4) per candidate level (A, C&E); some marking criteria have examples, especially the history taking one (e.g. 'Takes history of pain - onset, duration etc); there is some discrimination between A and C candidates.
3	Satisfactory number of marking criteria (4+) per candidate level (A, C & E); most marking criteria comes with examples/details of what the candidate is expected to do; there is a good level of discrimination between A and C candidates.
SP scripts	
1	Does not follow the current style guide; information is in long paragraphs and is hard to follow; does not respect the current guidance on order of information; it is relatively short and lacks details for the SP; some information is in the wrong sub-section.
2	Script follows most rules of the current guidance; information is structured more clearly but occasionally lacks detail (e.g. doses for drugs); most of the information is in the right order and correct section, with minor exceptions.
3	Script reflects the style guide closely, with minimal lapses; information is laid out in the expected order and has a good level of detail; the information is not misplaced in the wrong sections of the script; clear and

	concise.
Examiner's calibration guidance	
1	Is missing.
2	Generic guidance which could benefit from added comments but bespoke to the station.
3	An appropriate level of guidance that matches the needs of the station. Where generic, no more information is needed.
Examiner's supporting information	
1	Is missing or has been taken from sources generally best avoided (non-medical reference websites).
2	The information isn't clearly linked to the station or covers it only broadly, but it is sourced from acceptable websites; formatting could be improved and may be difficult to follow/find information quickly.
3	Guidance is sourced from NHS, CKS, NICE guidelines or recognisable medical websites from UK or abroad; the information is structured well (sub-headings and bolded text are used); the information links up directly with themes covered by the scenario.

Table 3: Scoring guidance for rating of station factors