Going beyond hawks and doves – measuring degrees of examiner misalignment in OSCEs

Author

Matt Homer, Schools of Education and Medicine, University of Leeds, LS2 9JT, UK

m.s.homer@leeds.ac.uk

Abstract

Background

Minimising examiner differences in scoring in OSCEs is key in supporting the validity of the assessment outcomes. However, the common classification of extreme examiners as ‘hawks’ or ‘doves’ can be overly simplistic. Rather, it is the difference in combined patterns of scoring/grading across OSCE circuits that better indicate poor levels of agreement between examiners - this misalignment can unfairly advantage particular groups of candidates in comparison with others in other circuits.

Methods and materials

In this study, a new measure of differences in examiner scoring is presented that quantifies the different combined patterns of scoring in global grades and station total scores for pairs of examiners assessing in the same station but in different circuits. Over 10,000 separate station administrations from 2016 to 2024 are analysed from the UK exam for international medical graduates who want to work in the NHS (PLAB2). The new misalignment measure is based on calculating the area between separate examiners’ individual borderline regression lines.

Results and conclusions

Particular station examples are presented where alignment between examiners is excellent and where it is poor. Longitudinal evidence suggests that average misalignment has declined over time suggesting that a range of interventions/developments of PLAB2 to improve calibration between examiners, and scoring practices more generally, have had some success. Variation in misalignment does not vary much between different types of station (e.g. standard/prescription/practical/simulation).

In challenging the ‘hawks’/‘doves’ paradigm, this paper contributes to the theoretical debate around the nature of examiner stringency in OSCEs. It also presents a new empirical misalignment measure which can be used to provide additional validity evidence for an OSCE-type assessment. Further work is needed to develop the metric to larger scale OSCEs.

Key words

OSCEs; metrics of quality; examiner stringency, ‘hawks’ and ‘doves’

Practices points

- Examiners can vary in their judgments of performance in OSCEs but classification as ‘hawks’ and ‘doves’ is overly simplistic in many contexts

- A new measure of differences in patterns of scoring across OSCE circuits by pairs of examiners is presented.
- This can be interpreted as a measure of misalignment between examiners and moves beyond the usual ‘hawks’ and ‘doves’ classification
- Evidence is presented as to how this measure can be used to improve fairness to candidates and to enhance the validity argument for OSCE-type assessments and associated outcomes.

Highlights

This paper develops a new way to understand and quantify differences in patterns of examiner scoring across parallel circuits in OSCE and shows how this measure can be used to improve OSCE validity and fairness to candidates.

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 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.

Introduction

OSCEs and other performance assessments are complex in nature and highly dependent on the quality of judgment of examiners of the candidate performance (Khan, Ramachandran, et al., 2013; Khan, Gaunt, et al., 2013; Harden et al., 2015; Yeates et al., 2019; Homer, 2022). In many contexts, there is an expectation that institutions will carry out detailed *post hoc* analysis of OSCE data in order to maximise their quality (Pell et al., 2010; Harden et al., 2015, chap. 14; General Medical Council, 2024a). Such analyses can aid decision-making in the short-term, for example providing justification for the removal of a poorly performing station from the assessment. In addition, measures of station and overall assessment quality can inform evaluation of interventions aimed at OSCE improvement over the longer term (Fuller et al., 2013).

However, the common classification of extreme examiners as either ‘hawks’ (stringent with marking) or ‘doves’ (more generous with marking) is overly simplistic in scenarios where candidate performance is scored in both a holistic and a more fine-grained way (i.e. via global grades and checklist/domain scoring respectively). It is the combined pattern of scoring that really matters. To reflect this broader understanding of examiner behaviour in such contexts, a new station-level OSCE metric is developed in this paper – one that is appropriate for quantifying the difference in patterns of scoring in the common situation in which each station is scored in two different ways - holistically via a global grade, and in a more fine-grained checklist or domain scoring approach – and where the borderline regression method (BRM) of standard setting is used (McKinley and Norcini, 2014; Harden et al., 2015, p.145). For pairs of examiners assessing in separate circuits, this new metric calculates the area between pairs of examiners’ individual borderline regression lines. This provides a robust measure of how closely their patterns of scoring align (i.e. how well calibrated their scoring is), and has a range of applications including comparing across stations to identify relatively poorly calibrated stations. It can also be used to assess the impact of longer-term examiner training interventions aimed at improving calibration across circuits within the OSCE

In short, this new metric rests on a more nuanced understanding of differences in examiner behaviour, and can contribute additional evidence to all stakeholders for the validity of the OSCE and its outcomes. In particular, it can help provide evidence to assuage fairness concerns that can arise in candidates comparing experiences across circuits.

The paper continues as follows: I give a comprehensive description of the new metric, and then exemplify its usage in a particular summative assessment context (PLAB2 in the UK). I conclude the paper with some comments on potential generalisations of the metric across multiple circuits, and other possible avenues for further research.

A new measure of differences in scoring patterns for pairs of examiners

In many OSCE contexts, it is very difficult to disentangle examiner and candidate effects as both are usually nested in parallel circuits (Swanson et al., 2013; Khan, Gaunt, et al., 2013; Yeates et al., 2019; Homer, 2022). A further level of complexity is that OSCE performances are often scored in two different ways – via a single global grade that provides a holistic overall measure of the quality of the candidate performance, and in a more structured way either via a checklist[1] or a set of domain scores that are usually totalled to provide a second measure of candidate performance in the station (Khan, Gaunt, et al., 2013). This complexity implies that simple measures of examiner stringency based on a single measure of candidate performance (i.e. only global grades or only total domain scores) cannot accurately capture the impact of different scoring patterns across examiners on candidates (Homer, 2024). In order to fully understand the impact of different patterns of examiner scoring, it is necessary to take into account both types of station scoring simultaneously in the analysis.

Consider, then, a single OSCE station that is administered across two parallel circuits (‘orange’ and ‘blue’) where, in each, the candidate is assessed by a single examiner via a global grade and a total domain score. In such a situation it is likely that borderline regression will be used to set the cut-score in the station using all data from both circuits/examiners to regress total domain scores on global grades within the station (McKinley and Norcini, 2014).

Two hypothetical versions of scatter plots of candidate global grades (x) versus total domain scores as a percentage (y) for a station are shown in Figure 1 and Figure 2. In each figure, each dot represents a particular candidate, and the two within-examiner borderline regression lines are shown to emphasise the pattern of difference by examiner in scoring in the station. The overall standard would be set using the combined data, and the borderline regression line for this pooled data (not shown) would lie between the two lines in each figure.

In Figure 1, the two separate lines cross within the scatter plot which highlights the fact that for lower overall global grades (*fail*, *borderline*) candidates in the blue circuit are being awarded higher total domain scores compared to those in the orange. At higher levels (*satisfactory*, *good*) the opposite is true. Who then is the ‘hawk’ and who the ‘dove’ in such a case?

[1] From now on I largely ignore the possibility of checklist scoring at the station level, but the entire argument in the paper is applicable to both domain and checklist scoring.

Contrast this with Figure 2, where the two within-examiner regression lines do not meet within the scatter plot, and candidates within the blue circuit are advantaged in terms of domain scores across the full range of global grades. Is the blue circuit examiner to automatically be considered a ‘dove’ relative to the orange? Is it really that simple?

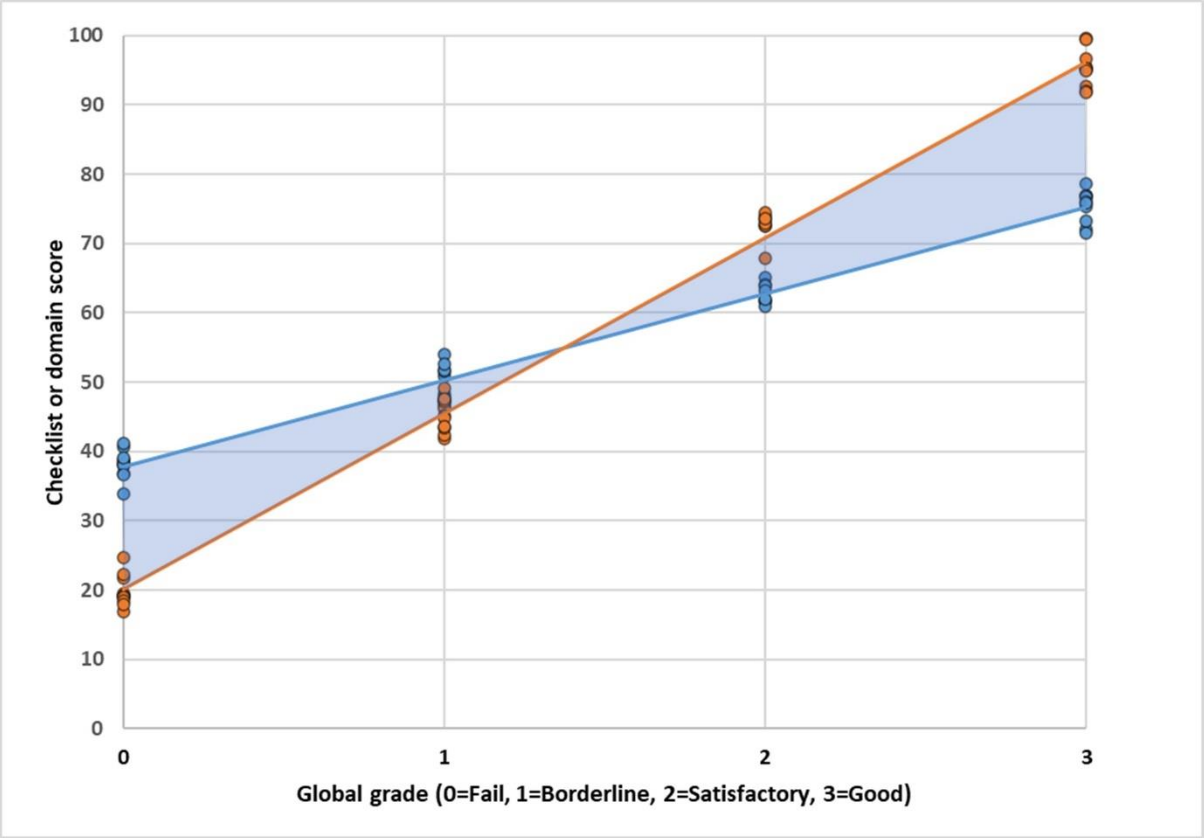

Figure 1: Scatter plot of grades versus domain score and within-examiner BRM lines that cross

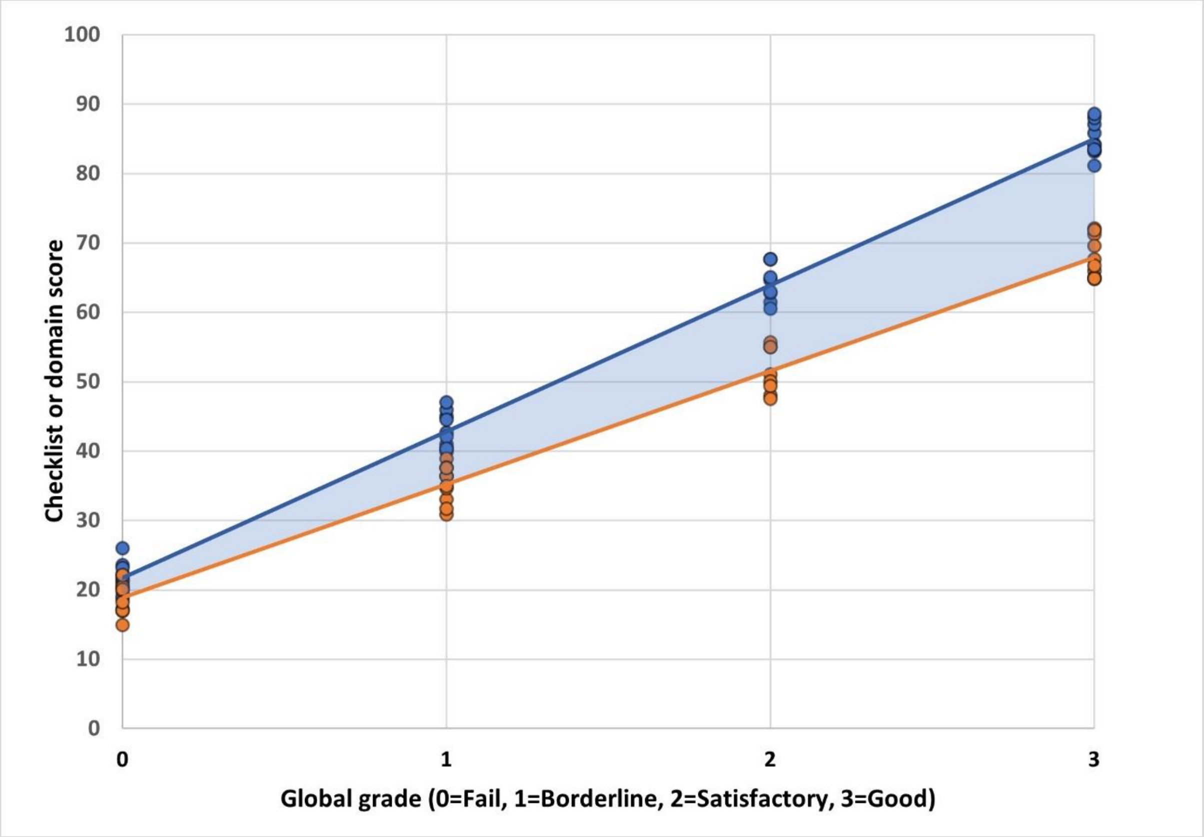

Figure 2: Scatter plot of grades versus domain score and within-examiner BRM lines that do not cross

These figures are archetypes of typical potential differences in scoring patterns across two parallel administrations of a station that uses both global grades and total domain scores within the station. The shaded area in each case provides a meaningful measure of these different patterns of scoring, derived using all global grades and all total domain scores within the station. These examples also illustrate how the 'hawk'/'dove' classification quickly becomes quite problematic.

An area of zero between the two lines would indicate that there was no difference in patterns of scoring, and the larger the area becomes the more variation across the two examiners in their pattern of scoring there is. To make this area metric more applicable to a range of contexts, it is best expressed as a percentage of the total area of the scatter plot. This facilitates comparisons across contexts where patterns of scoring or grading are different (for example, where there are five global grades rather than four).

Given the slope and intercept of each within-examiner BRM line, the area can be calculated using relatively simple algebra – the formulae corresponding to each figure are shown in the Appendix. The paper continues with a detailed empirical illustration of this new area metric from a particular OSCE context and concludes with a discussion of how the metric might be used and developed in the future.

Exemplification of the new metric in a particular context

The PLAB2 exam

The data in this study comes from PLAB2 - the summative clinical assessment in the UK for international medical graduates who want to work in the National Health Service (General Medical Council, 2024b). In its current post-COVID format, it is 16 station OSCE set at a level to ensure that successful candidates can provide good care equivalent to those UK-based trainee clinicians in their second year of post-graduate training (called FY2 in the UK).

Stations are 8 minutes long with 1.5 minutes of reading time between stations. A single trained and clinically qualified examiner is present in each station. Depending on the number of candidates for each exam, there are either one or two circuits, and the examiner is typically present in the same station for two sessions so that they usually 'see' of the order of 30 candidates in each exam. On the morning of the exam, station-level calibration takes place between pairs of examiners and simulated patients across parallel circuits to maximise consistency across stations.

Given the relatively strong international demand for PLAB2, the exam takes place on most working days throughout the year. For example, in 2023 there were 275 separate PLAB2 exams with over 4,300 station administrations and a median of 62 candidate per exam.

A group of between 20-30 senior clinicians and other assessment experts meet bi-monthly to oversee the PLAB2 assessment practice in terms of providing input into quality assurance issues, implementing changes to medical practice, writing and developing stations, and discussing other longer-term change and improvement to PLAB2.

In terms of scoring, the candidate performance in each station is assessed via a single global grade on a scale from 0 to 3 [2] providing an overall holistic judgment of candidate performance in the station, and by a total domain score on a scale from 0 to 12 [3].

At the station level, the format of the PLAB2 exam has been stable since November 2016. Since that time, station-level pass/fail decisions have been made using Borderline regression (Kramer et al., 2003; McKinley and Norcini, 2014) with a median pass rate of approximately 69% across all stations.

Data across parallel circuits

Over the period November 2016 to October 2024 there were 10,226 separate station administrations with exactly two examiners in parallel circuits – approximately 46% of all station administrations over this period[4]. These two-circuit administrations form the main dataset for this study with separate values for station slope and intercept available for each of the two examiners assessing in parallel. This allows the use of the formulae in the Appendix to calculate the area metric for each of these >10,000 station administrations.

The empirical distribution of the area metric

Figure 3 shows the histogram for the area (percentage) metric in the study (n=10,226, median=5.70, mean=6.55, 5th percentile=1.39, 95th percentile=14.79). The distribution is positively skewed, and the most problematic stations are those that are to the right of the distribution where the larger values correspond to greater misalignment between pairs of examiners in their patterns of marking across the two circuits.

[2] 0 = *fail*: could not carry out work of a day one FY2, 1 = *borderline*: not convinced could carry out work of a day one FY2, 2 = *satisfactory*: could carry out work of a day one FY2 safely, 3 = *good*: could be expected to carry out work of a day one FY2 to a high standard

[3] Each of three domains is scored 0 to 4 (*Data gathering, technical and assessment skills*; *Clinical management skills*; and *Interpersonal skills*) and there are positive and negative station-specific key feature descriptors to guide examiners in their judgments in each domain.

[4] 22,480 in total with the remainder (54%) with only a single examiner.

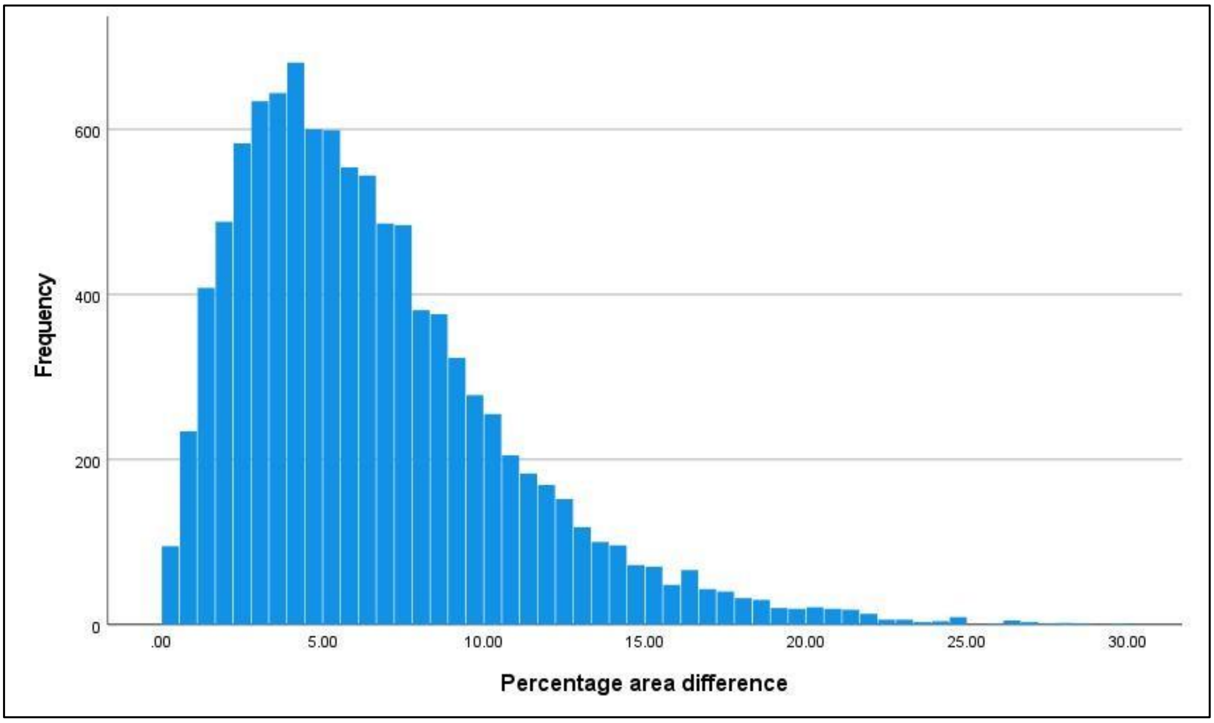

Figure 3: Histogram of percentage area between regression lines for stations with two examiners

The analysis shows that in 56% of these station administrations, the two regression lines meet within the scatter plot (as in Figure 1) whereas the remaining 44% the lines do not (as per Figure 2).

To illustrate the potential use of the new metric, three particular station administrations are now discussed alongside the corresponding scatter plot of grades and total domain scores. The final part of the empirical work investigates changes in the metric over the period of the study and quantifies differences in the area metric across different types of station.

Removing a station

The station shown in Figure 4 was removed from the exam based on the standard post-exam inspection process that takes place soon after the completion of the exam. However, it would also have been automatically identified based on its large area metric (area=22.6%[5]). This station is an extreme version of that shown in Figure 2. In the current case, candidates in the green circuit are very likely being advantaged compared to those in the blue since, at the same level of apparent overall performance (global grade), they outscore them by of the order of two domain score marks (out of 12). Without its removal, there would be nine candidates from the blue circuit failing the station who were awarded *satisfactory* grades, whereas all candidates awarded this grade in the green circuit would pass. Of course, we cannot be sure *post hoc* what has gone wrong in this station in terms of scoring, but what we can say is that based on the scatter plot and/or the area metric, the degree of poor alignment in scoring/grading between examiners is a genuine problem, and fairness considerations to candidates imply that it is best removed from the examination. In fact, the removal of a station is quite a rare occurrence in PLAB2 (of the order of 1% of stations or less).

[5] Note that the area metric is calculated using the full range of scoring, even in cases where, perhaps, there are no candidates present (e.g. for the blue circuit in Figure 4 at the *good* grade).

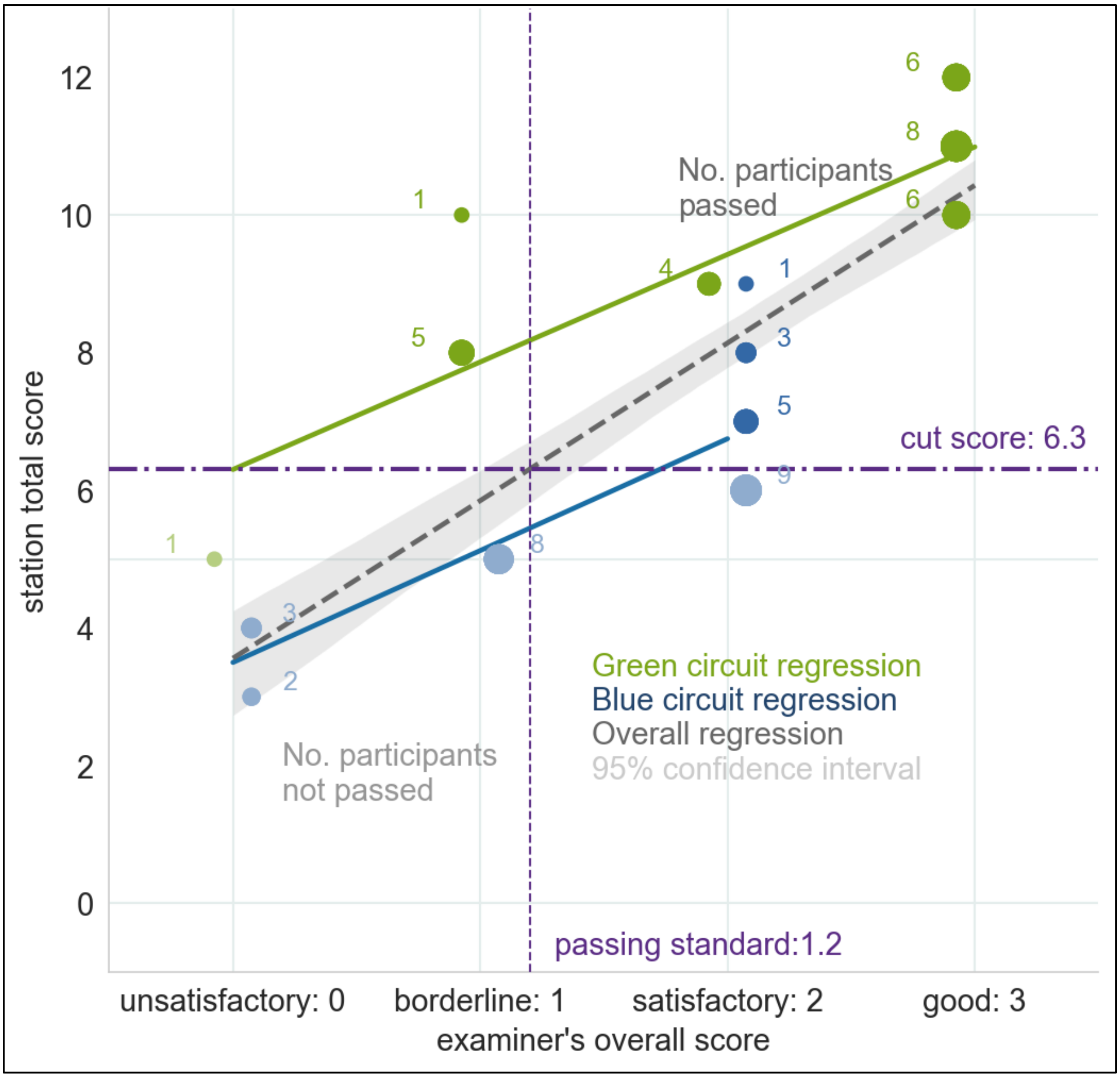

Figure 4: Scatter plot of a removed station with a large area metric

As part of the usual process when a station has to be removed from PLAB2, post-exam feedback was provided to both examiners in the form the scatter plot itself.

A closely aligned station

In Figure 5, there is very close alignment in the two regression lines across the two circuits despite the actual scoring/grading being somewhat different across circuits. The overlapping of the two lines is reflected in the small area metric for this station (area metric=0.11%).

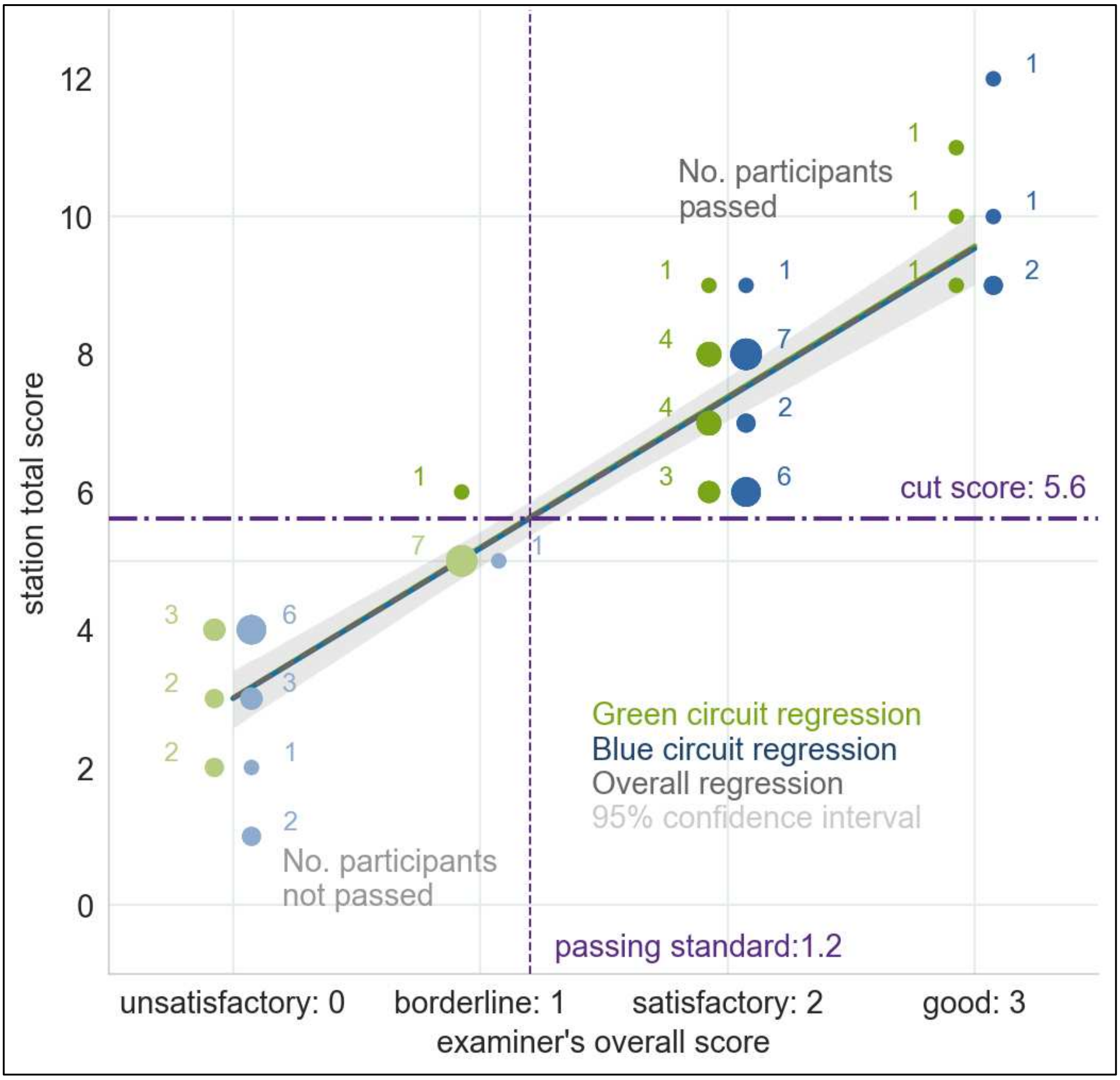

Figure 5: Scatter plot of a station with close alignment between examiners

A station where the area metric can be misleading

The area metric for the scatter plot in Figure 6 has value 27.5% - one of the largest in the whole dataset. However, this is largely an artefact of the lack of spread in scoring which means that the regression lines are extrapolated away from the actual scoring when the area is calculated. In this administration, almost all candidates performed poorly with only one (out of 61) passing the station. This example underlines the importance of always inspecting scatter plots of stations and not just relying on summary metrics to make decisions about station performance.

Historic data on this station suggests that the area metric is usually closer to the overall average across all stations where two examiners are present (median area=6.1% on 60 administrations, overall median across all stations=5.7%). The data also shows that this station does not usually have such a high failure rate. Hence, the overall implication here is that there is no problem with this station (or with examiners), but rather that the large area metric is a result of weak cohort performance in this particular administration.

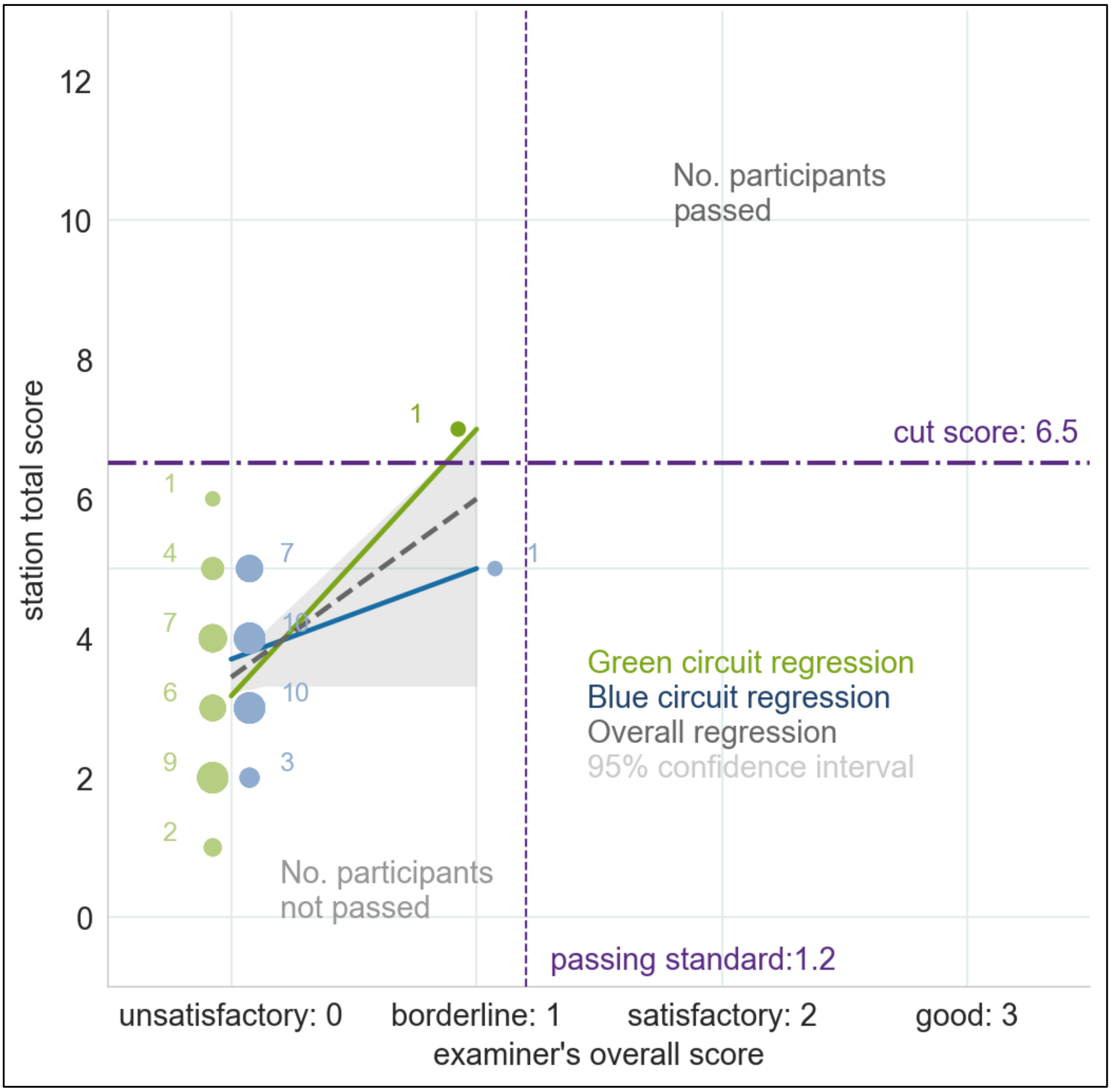

Figure 6: Scatter plot of a station with large area metric based on extrapolation away from the data

Changes over time and differences across station types

A Pearson correlation of date of exam versus area metric in the PLAB2 data gives a statistically significant negative correlation (r=-0.24, 95% bootstrapped confidence interval [-0.25, -0.21], n=10,226, p<0.001). This indicates that there has been an important decline in the typical size of the area metric over the period of the study. In turn, this suggests that the ongoing quality control efforts within the PLAB2 exam administration and governance processes have had some success in decreasing unwanted differences in patterns of scoring across examiners in parallel circuits. These interventions include improved examiner training, better calibration practices on the day of the exam with a dedicated calibration checklist, more focussed and specific scoring descriptors in stations, better post-exam feedback to problematic examiners, and an improved annual examiner appraisal system.

One might expect that different types of stations might be more or less easy to successfully calibrate. As part of the blueprinting process, PLAB2 stations are classified into one of five categories (*Standard*, *Practical*, *Telephone*, *Prescription* and *METI* – a high-fidelity simulator) based on the key activities taking place in the encounter[6]. A one-way ANOVA shows that whilst there are statistically significant differences in mean area across station types, the overall effect size is quite small (F(4,9524)=8.71, p<0.001, R-squared=0.004). The only statistically significant Bonferroni *post hoc* test was for *Standard* stations compared to each of the other four levels of station type. Figure 7 shows the corresponding error bar for this analysis and indicates that *Standard* stations (which make up around two-thirds of all stations), have the highest typical area metric. In other words, these stations have the greatest misalignment problem across circuits, compared to other station types.

Figure 7 also indicates that the *METI* stations have the lowest mean area metric – suggesting that these stations might be marginally easier in terms of examiner alignment of scoring compared to other types of stations. However, these differences are quite small, and the overall results here do not suggest that the nature of the classification of a station to a particular type strongly impact on calibration issues within stations. This provides some re-assurance to OSCE developers/writers and other stakeholders that any concerns of misalignment in examiner scoring is not driven (much) by station type.

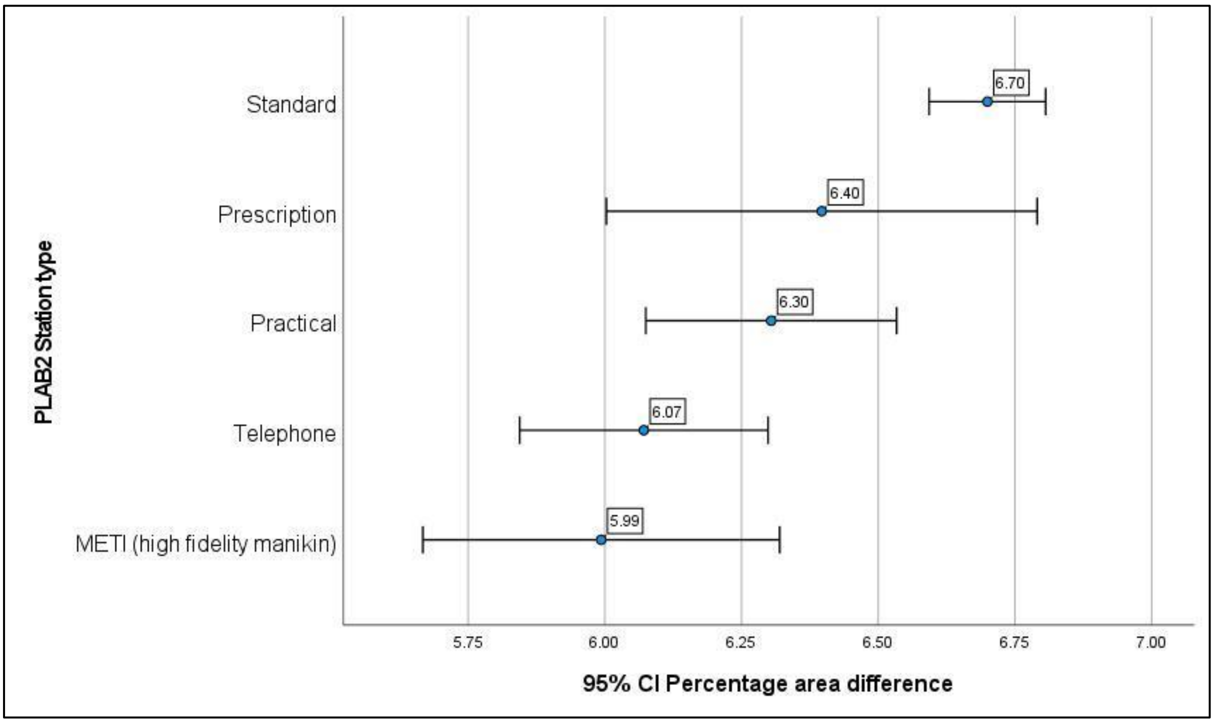

Figure 7: Error bar (mean and 95% confidence interval) for area metric by station type

Concluding remarks

This study emerged from a critique of the standard notions regarding extremes of OSCE examiner behaviour (‘hawks’ and ‘doves’). As a result, a new way to quantify differences in

[6] 7% of stations in the data were missing this classification for administrative reasons.

patterns of examiner scoring across two parallel circuits in OSCEs has been developed and investigated. The new area metric uses all scoring in the station to do this (i.e. both global grades and total domain scores). The metric has clear application as a *post hoc* station-level metric – for example, it can be used to reveal where calibration in stations has not worked well (or vice versa), and to monitor over time the impact of interventions aimed at improving calibration practices across circuits (Pell et al., 2010; Fuller et al., 2013). In the end, evidence provided by this metric can contribute to the fairness of the assessment and to its broader validity.

The metric has already become a new addition to the regular set of ongoing *post hoc* quality control reporting in PLAB2 (Pell et al., 2010). It has the advantage that it provides a robust comparison across circuits, and is not negatively impacted by the nesting of candidates and examiners in these in the way that, for example, the calculation of reliability/reproducibility coefficients is (Swanson et al., 2013; Homer, 2022). Over time and with sufficient data it can be used to identify examiners who tend to be more misaligned than others. As with most OSCE-type analyses, the methods presented in this paper could also be applied to help quality assure admissions data (e.g. MMIs) where there are parallel circuits (Cleland et al., 2023).

It is important to recognise that in many institutions (for example large medical schools) the number of parallel circuits in an OSCE is much larger than two - i.e. of the order of 20 or more (Khan, Gaunt, et al., 2013; Harden et al., 2015) - so the metric would need to be generalised to work in these contexts provided there is enough data points for each examiner. One way to do this might be to carry out the analysis as presented for all combinations of each pair of examiners in the station. With 20 parallel circuits this would give 190 (= C_2^{20}) area values for each station administration so this work would clearly need to be automated in some way (e.g. using R or Python scripts). Further statistical analysis would then be needed to interpret this distribution of area values, probably comparing across stations via measures of area location (mean) and spread (standard deviation) and to investigate, for example, whether there are individual examiners who look like outliers compared to their peers within the station. There is a clearly an additional level of complexity in generalising the approach to beyond that present in the PLAB2 context. but future work could certainly provide additional insights into differential examiner marking behaviours within and across stations in a multi-circuit OSCE.

Another area of applicability for this work would be in a comparison across different sites in a multi-site OSCE (Harden et al., 2015, p.85). In other words, the unit of analysis for the area metric would be the site, not the examiner. All data for each site could be pooled and then a comparison made across these different sites to investigate systematic differences in patterns of scoring across these for each station in the exam. Given what is already known about the relatively large differences in examiner behaviours across sites (Sebok et al., 2015; Yeates et al., 2024), this could provide informative quality assurance data in such designs. It might also be possible to extend such work to include site and examiner effects simultaneously, but this would need additional development.

In terms of study limitations, this work should be regarded as foundational rather than definitive given that it covers in detail only the case of two parallel circuits, and from a single assessment context. There would obviously be a major benefit in the production of complementary studies from other contexts. This would allow both inter- and intra-study comparisons and help situate the findings from this study in a wider set of literature.

Ethics, funding and data access

PLAB candidates consent to their examination data being used for research purposes. 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 author acts as a paid assessment consultant to the PLAB2 panel on an *ad hoc* basis. For reasons of confidentiality and assessment security, the data used in this study is not available to other researchers.

Declaration of interest

I have no declarations of interest to report.

Acknowledgments

The author would like to acknowledge the help and support of the PLAB2 team at the UK General Medical Council in aiding the early development of this work.

References

Cleland, J., Blitz, J., Cleutjens, K.B.J.M., oude Egbrink, M.G.A., Schreurs, S. and Patterson, F. 2023. Robust, defensible, and fair: The AMEE guide to selection into medical school: AMEE Guide No. 153. *Medical Teacher*. **45**(10), pp.1071–1084.

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 2024a. Clinical and professional skills assessment (CPSA). *Clinical and professional skills assessment (CPSA)*. [Online]. [Accessed 21 October 2024]. Available from: https://www.gmc-uk.org/education/medical-licensing-assessment/uk-medical-schools-guide-to-the-mla/clinical-and-professional-skills-assessment-cpsa.

General Medical Council 2024b. PLAB 2 guide. [Accessed 12 November 2024]. Available from: https://www.gmc-uk.org/registration-and-licensing/join-the-register/plab/plab-2-guide.

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.

Homer, M. 2022. Pass/fail decisions and standards: the impact of differential examiner stringency on OSCE outcomes. *Advances in Health Sciences Education*. **27**(2), pp.457–473.

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

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.

Khan, K.Z., Ramachandran, S., Gaunt, K. and Pushkar, P. 2013. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. *Medical Teacher*. **35**(9), pp.e1437-1446.

Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L. and van der Vleuten, C. 2003. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. *Medical Education*. **37**(2), pp.132–139.

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.

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

Sebok, S.S., Roy, M., Klinger, D.A. and De Champlain, A.F. 2015. Examiners and content and site: Oh My! A national organization’s investigation of score variation in large-scale performance assessments. *Advances in Health Sciences Education*. **20**(3), pp.581–594.

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., Cope, N., Hawarden, A., Bradshaw, H., McCray, G. and Homer, M. 2019. Developing a video-based method to compare and adjust examiner effects in fully nested OSCEs. *Medical Education*. **53**(3), pp.250–263.

Yeates, P., Maluf, A., McCray, G., Kinston, R., Cope, N., Cullen, K., O’Neill, V., Cole, A., Chung, C., Goodfellow, R., Vallender, R., Ensaff, S., Goddard-Fuller, R. and McKinley, R. 2024. Inter-school variations in the standard of examiners’ graduation-level OSCE judgements. *Medical Teacher*. **0**(0), pp.1–9.

Appendix

Suppose the maximum achievable domain score is D (=12 in PLAB2) and the global grade is on a scale 0 to G (=3 in PLAB2).

Also, assume the two regression lines have slopes $m1, m2$ and intercept $c1$ and $c2$ respectively.

Assuming $m1 \neq m2$, the two lines cross at an x-value given by:

$$I = (c2 - c1)/(m1 - m2)$$

The formula for the area is most easily calculated using integral calculus to find the area between two lines between appropriate x-values.

There are two distinct cases:

Regression lines that cross within the scatter plot (as per Figure 1)

The area is given by the absolute value of:

$$A = \frac{(m2 - m1)G^2}{2} + (c2 - c1)(G - I)$$

Regression lines that don't cross within the scatter plot (as per Figure 2)

The area is given by the absolute value of:

$$A = \frac{(m2 - m1)G^2}{2} + (c2 - c1)G$$

To calculate the percentage area in either case multiple by 100 and divide by DG.

An R script calculating the area is available on request from the author. This script first checks whether the two lines meet within the scatter plot, and then branches to the appropriate formula as given above.