Influences on scoring and grading in PLAB2:

PLAB2 Annual Report 2022 - December 2023

Dr Matt Homer, Leeds Institute of Medical Education, University of Leeds

m.s.homer@leeds.ac.uk

Contents

List of figures

List of tables

Executive summary

Previous research indicates that there is a need to better understand factors influencing station-level scoring and grading in PLAB2.

This report therefore investigates the association between a range of examiner characteristics (e.g. examiner sex, examiner ethnicity…) and station-level PLAB2 outcomes (i.e. candidate-level total domain scores and global grades). It also includes the influence on scores and grades of station type, and that of other important factors at the station level (e.g. the candidate, examiner, station, and exam involved).

The data is from March 2022 to October 2022 across 290 PLAB2 exams, with a total of 189,862 rows of candidate scores/grades in stations for most analyses, although for some analysis there is a small proportion of missing data on some variables.

Simple (single predictor) and multivariate modelling of total domain scores and, separately, grades in stations are carried out across all characteristics and factors.

Key findings

Individual influences on scores and grades

In a simple (i.e. single predictor) model for PLAB2 scores and grades:

- Most examiner characteristics have little overall influence on scores and grades (overall summary in Table 4).
- However, some categories of some characteristics might show some arguably important differences (e.g. for *Examiner ethnicity* see Figure 6, Figure 7).
- *Station type* does show clear differences in average scoring and grading across categories of this variable – with *Prescription* stations scoring/grading the lowest and *Standard* stations the highest (Figure 22, Figure 23).
- *Examiner Primary medical qualification country of origin* shows quite large differences in average scoring/grading across countries (Figure 20, Figure 21).
- The effects of *Candidate*, *Examiner* and *Station* (treated as random effects) typically have much stronger predictive power on scoring/grading than do other examiner characteristics or station type (summary in Table 5). *Exam* has little effect.

Multiple influences on scores and grades

In a combined (i.e. multivariate) model for PLAB2 scores and grades using a range of examiner and station characteristics, and other random effects (*Candidate*, *Examiner*, *Station*…):

- The random effects of *Examiner*, *Candidate* and *Station* are most important in influencing total domain scores, and the findings are similar for global grades (Table 7).
- The influence of *Exam* and *Examiner PMQ country of origin* are marginal once other factors are included in the model (Table 7).
- *Station type* has a statistically significant role in influencing scores and, consistent with the simpler analysis, *Prescription* stations are scored and graded the lowest on average (Table 8).

- The only other statistically significant effect is for *Hindu* examiners whose scoring/grading is on average lower than that of examiners whose religion is *Christian* (the reference group for that variable; Table 8).

Conclusions

The main conclusion of this work is that few of the factors included in the analysis have been found to be important in influencing PLAB2 station-level outcomes (whether scores or grades). Almost all examiner characteristics are relatively unimportant in this regard, and only station type, as might have been expected in advance, shows some important differences.

Those factors that do show important influences are examiner, candidate and station – in ways that are similar to that seen in previous work on PLAB2 and in other OSCE-type assessments. Whilst the 'error' due to differential examiner stringency is, to an extent, a problem at the station-level, the literature suggests that at the exam level these differences largely cancel each other out. OSCE-type assessments require an ongoing effort to maintain quality and validity, and this work confirms that examiners are one of the key factors that threaten this.

There is little in the way of policy recommendations for PLAB2 to be made as a result of this report. However, the variation in scoring/grading across stations underlines the need to move to a more defensible minimum station hurdle to take some account of the mix of station difficulties in each exam. Pilot work on this is currently in progress.

Introduction

A range of recent work on PLAB2[1] and the wider literature, suggests that examiners vary in how they score OSCE performance (Yeates and Sebok-Syer, 2017; Homer, 2022; Homer, 2023b). This is the case despite a range of quality control measures in place designed to maximise standardisation of scoring - including for example, extensive examiner training and monitoring, and on-the-day calibration processes.

In an effort to better understand what might be driving differences in scoring/grading, this report takes a range of examiner characteristics and investigates their association with PLAB2 outcomes at the station-level – both in simple (single predictor) analysis, but also in a combined, more complex multivariate modelling approach. This work contributes to the literature in a key area of interest in health professional educational assessment, using rich PLAB2 datasets that are unusual in that they allow for the disentangling of a range of important assessment factors (i.e. candidate, examiner, and station). It might also give some suggestions for the future development of PLAB2.

The key examiner characteristics in the analysis include sex, ethnicity, and details of their training (Primary Medical Qualification country of origin, and date of first registration). Also included is a variable that classifies stations by their type according to a PLAB2 schema (e.g. standard stations, Skype/telephone stations and so on). Details of all variables are given in the methodology section that follows.

In the remainder of this report we describe briefly the methodologies employed, then present the findings and make some concluding remarks.

Methodology

PLAB2 candidate-level total station domain scores and examiner overall judgments from March 2022 to October 2022 December were merged with examiner and station level data from the same period. This recent period was chosen as the exam was in a stable format over the whole of that time.

The original dataset consisted of 203,868 rows of candidate scores/grades. After data cleaning and merging 189,862 rows of complete data (93.1%) were employed in the main analysis. For a range of reasons, there were some records where the examiner code was not captured, and other records where scores/grades were missing from stations suppressed in the exam.

Table 1 shows the sample sizes of various facets (factors) of the PLAB2 data used. These will be treated as random effects in the modelling – where they are considered as samples from a wider potential population.

[1] Under arrangements for the forthcoming UK Medical Licensing Assessment, PLAB2 will become referred to as the CPSA (Clinical and Professional Skills Assessment) in 2024. In this report, we keep with the PLAB2 nomenclature that was used during the generation of the data employed (from 2022).

Exam facet	Sample size (i.e. number of distinct levels of each facet)	Median occurrence of each level in the data	Description
Exam	290	496	There are 290 different exams in the dataset, and typically there are 496 rows of data for each exam.
Examiner	840	123	There are 840 different examiners in the dataset, and typically each is present in 123 rows of data.
Station	533	299	There are 533 different stations in the dataset, and typically each is present in 299 rows of data.
Candidate[2]	12,749	15	There are 12,749 different candidates in the dataset, and typically each is present in 15 rows of data.

Table 1: Typical levels of each facet in analysis

The first station-level outcome measures that we will investigate is candidate total domain score (usually referred to as 'score' in what follows). This is on a scale from 0 to 12 based on the summation across the three content domains that PLAB2 assesses in each station (*1. Data gathering, technical and assessment skills; 2. Clinical management skills; 3. Interpersonal skills*).

The second PLAB2 outcome is the overall examiner judgment of candidate performance in each station (referred to as a global grade in what follows).This is scored 0=*fail*, 1=*borderline*, 2=*satisfactory*, 3=*good* - with reference to performance as a day one Foundation Year 2 doctor.

Differences in these outcomes across a range of examiner characteristics will be investigated (*Sex*, *ethnicity*, *disability*, *Sexual orientation*, *religion*, *Date of first registration*[3], *Status as GP or not*, *Status as a specialist or not*, and *Country of Primary medical qualification*).

In PLAB2 all stations are characterised as either *Standard*, *Skype/Telephone*, *Practical*, *METI* or *Prescription*. This is another variable we will investigate in terms of its influence on PLAB2 outcomes.

---

[2] Given that there is good evidence that PLAB2 candidates performance can change over successive attempts (Homer, 2022), each separate candidate attempt within this period was treated as if from a new candidate.

[3] Also available in the data were *(*examiner*) Age* and *Date of passing primary medical qualification*. However, these were strongly correlated with *Date of first registration* so only this latter variable is included in the analysis. *Date of first registration* is the only scale predictor in the dataset that we use.

Univariate modelling of scores and grades with a single predictor

In terms of methods of analysis, as simple methods as possible are used – initially including purely descriptive and graphical representations. The General Linear Model (GLM), essentially a generalisation of multiple regression that allows for categorical or scale predictors (Field, 2013, chaps 11–12), is used to compare scores (and separately grades) across groups/variables. This is mainly a set of relatively simple analyses investigating the extent to which scores, and separately grades, vary across each level (category) of a single predictor.

For interpreting results, the focus is on appropriate effect size measures (mainly R-squared for GLM) rather than p-values (Amrhein et al., 2019; Wasserstein et al., 2019). This is important given the relatively large overall sample sizes in the data (Table 1) which can produce statistically significant results on relatively small, not very important, observed differences. Also, the p-values (which indicate which variables are playing a statistically significant role) are likely to be under-estimated given the interdependence in the data that these GLM analyses ignore – for example, examiners, candidates and stations repeat across many rows of the data (Hutchison and Schagen, 2008).

R-squared as the effect size has a relatively simple interpretation as a measure of how much variation in the outcome (i.e. score or grade) is captured by the predictor. This effect size measure is produced regardless of the nature of the predictor (i.e. as either scale or categorical) and so by comparing across analyses we can see which single predictors have the strongest effect on the outcome scores or grades. As well as R-squared, where appropriate we give additional descriptive information – for example, means across groups and associated error bars or scatter plots.

These analyses are carried out using SPSS (IBM Corp, 2021). Note that there are a few percent missing on some variables (e.g. examiner sex, ethnicity,…) where examiners did not respond or categories have very low numbers. We detail this missingness at appropriate points in the Findings section.

For the sake of completeness, we also produce a ‘null’ model for each facet from Table 1. This includes only the single facet, treated as a random effect, as a predictor and gives additional insight into which facets are most important in influencing PLAB2 outcomes.

More complex multivariate modelling of scores and grades

To investigate multiple influences on scores (and then grades) a combined (multivariate) analysis using mixed models is carried out using the R package lme4 (Bates et al., 2015). In this multivariate modelling with total domain score (or, separately, global grade) as the outcome, we treat most factors as fixed effects (e.g. *Examiner sex*, *Examiner ethnicity*, *Station type*,…) but *Candidate*, *Examiner*, *Station*, *Exam* and *Primary Medical Qualification (PMQ) country of origin* as a random (intercept) effect – given how many levels each of these variables have (see Table 1).

This modelling approach gives individual estimates for each level of the categorical fixed effects (e.g. *Examiner sex*, *Examiner ethnicity*, *Station type*,…) – relative to an appropriate single reference group. For the sole scale fixed effect, *Date of first registration*, it gives a single estimate (essentially a slope) indicating how the outcome varies with the predictor. All these estimates are interpreted in the combined model as having controlled for other factors in the modelling.

The mixed modelling also parcels out the remaining variance across the random effects included in the model and can also give an estimate for each level of each random effect. For example, for examiners this would be an estimate of their typical scoring (i.e. stringency) having controlled for all the factors present in the model.

Findings

We begin with aggregate information on the PLAB2 outcomes, and then present the simple analyses (with a single predictor), and end with the more complex mixed modelling.

Distributions of PLAB2 scores and grades

Figure 1 and Figure 2 show histograms of the distributions of scores and grades respectively across the full dataset. There is a good spread in both distributions, and for scores in particular there is relatively little skew. For completeness, we also include summary statistics for score and grades in Table 2.

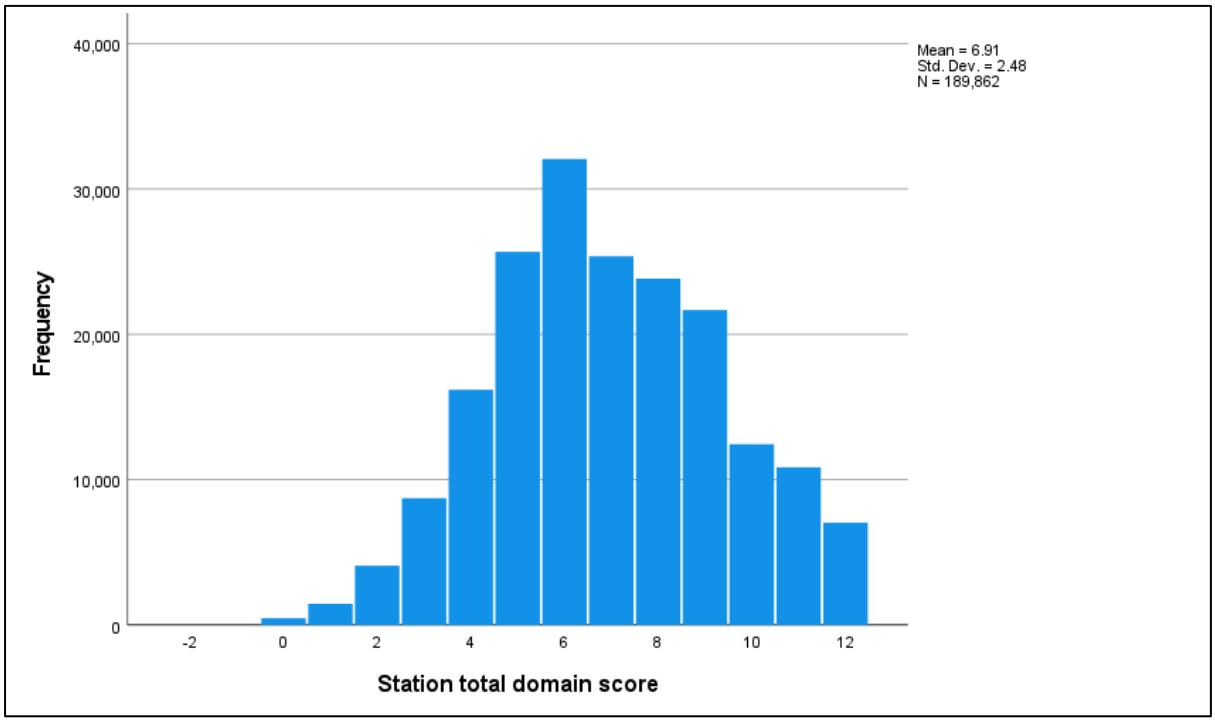

Figure 1: Histogram of station total domain scores

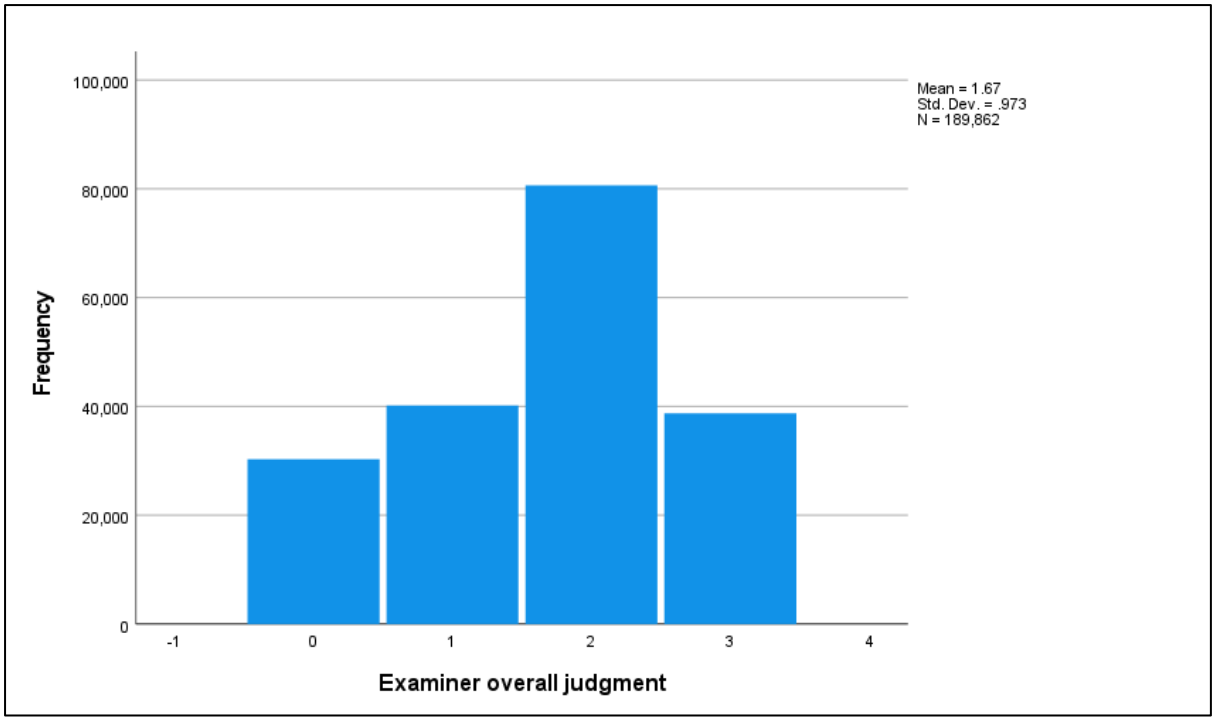

Figure 2: Histogram of station global grades

	Mean	Median	Std. Deviation	Skewness	Kurtosis	Quartiles		
Total domain score	6.91	7	2.48	0.065	-0.479	5	7	9
Global Grade	1.67	2	0.97	-0.348	-0.847	1	2	2

Table 2: Summary statistics for station scores and grades

The correlation (attenuated – i.e. not corrected for measurement error) between scores and grades is r=0.88 (n=189,862, p<0.001). Figure 3 shows the scatter plot of the relationship between the scores and grades given in each station.

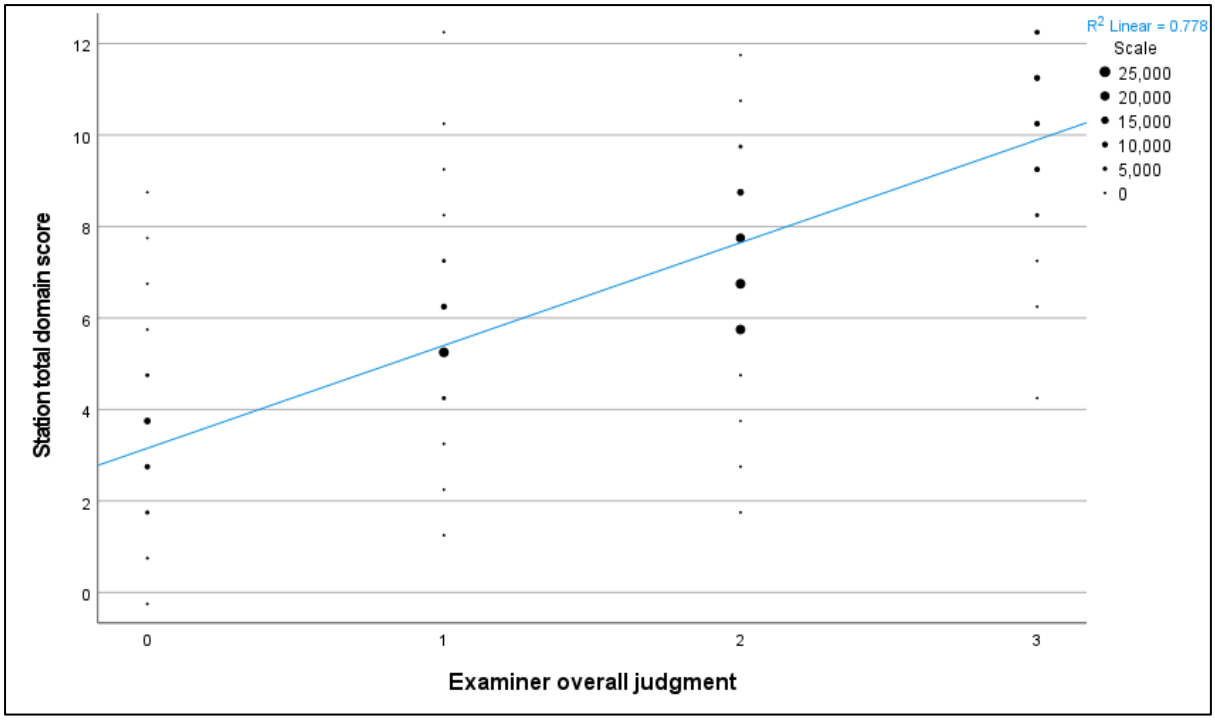

Figure 3: Scatter plot of total domain scores versus global grades

Individual influences on scores and grades – fixed effects

In what follows, each sub-section details simple single predictor analysis of PLAB2 scores and grades against examiner and station characteristics.

Examiner sex

Figure 4 shows that there is almost no difference in mean scores by *Examiner sex* (F(1,180233)=0.13, p=0.72, R squared=0.000).

Similarly, Figure 5 shows the same to be true for global grades (F(1,180233)=2.73, p=0.09, R squared=0.000).

5.1% of responses on this variable were missing.

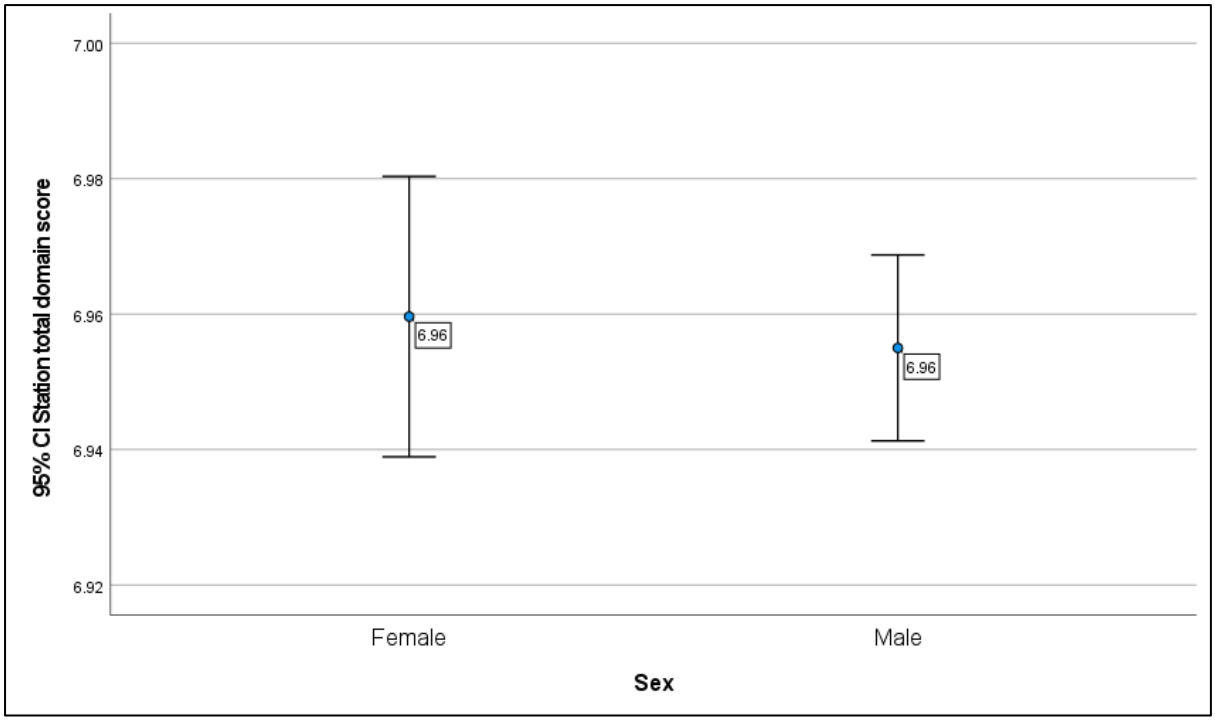

Figure 4: Error bar of total domain scores by examiner sex

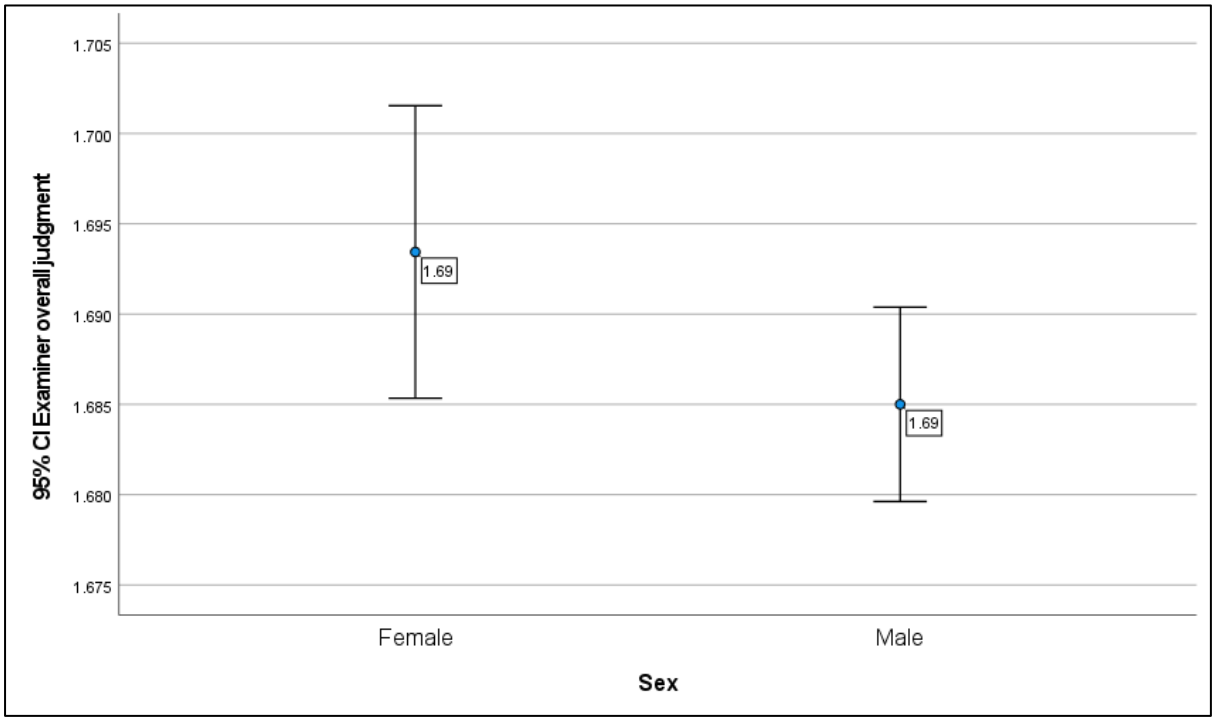

Figure 5: Error bar of global grades by examiner sex

Examiner ethnicity

There are some differences in scores and grades by Examiner ethnicity, but the overall effect is quite small.

Figure 6 (F(10, 189851)=112.5, p<0.001, R squared=0.006), and for global grades Figure 7 (F(10, 189851)=87.2, p<0.001, R squared=0.005).

There is some consistency across the two analyses – for example the *Asian / Asian British – Chinese* group appear more hawkish than other ethnic groups on both scores and grades.

2.0% of responses were missing on this variable.

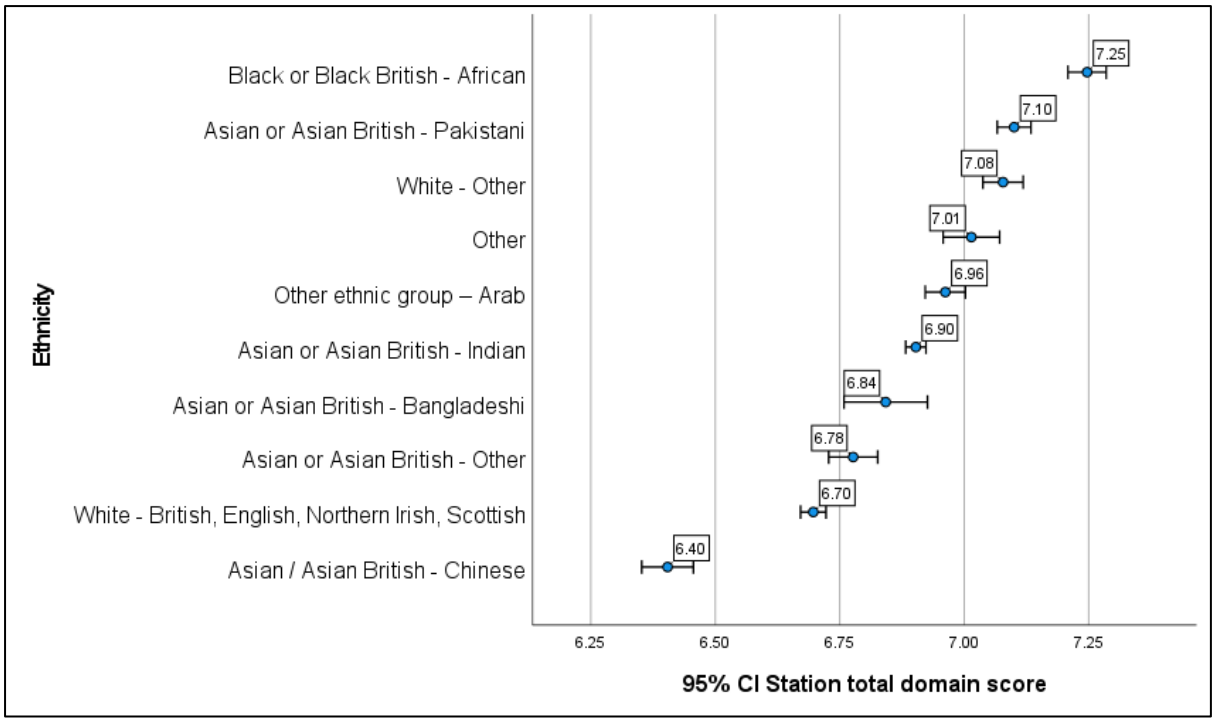

Figure 6: Error bar of total domain scores by examiner ethnicity

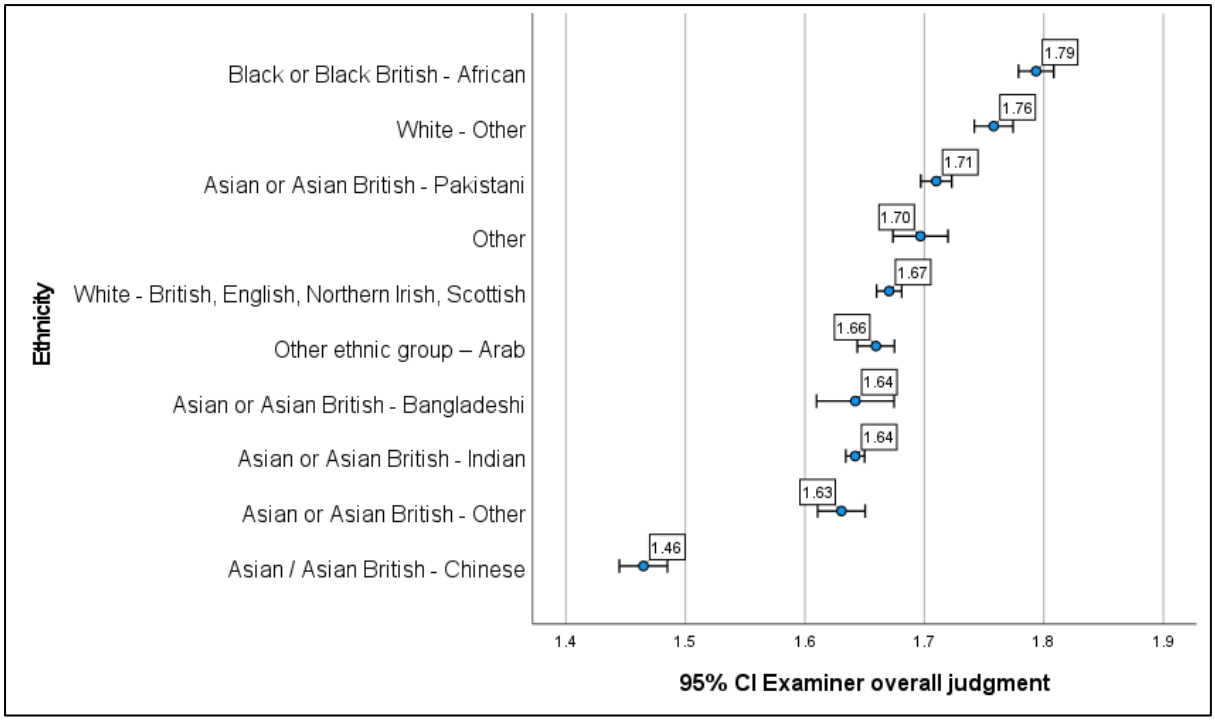

Figure 7: Error bar of global grades by examiner ethnicity

Examiner status as disabled

There are some small differences in scores and grades by *Examiner disability*, with non-disabled examiners tending to give higher scores and grades – see respectively Figure 8 ($F(1,178924)=14.3$, $p<0.001$, R squared=0.000) for scores, and Figure 9 $F(1,178924)=7.8$, $p=.005$, R squared=0.000) for global grades.

However, the disabled group forms only 1.5% of the dataset, and 5.8% of responses were missing on this variable overall.

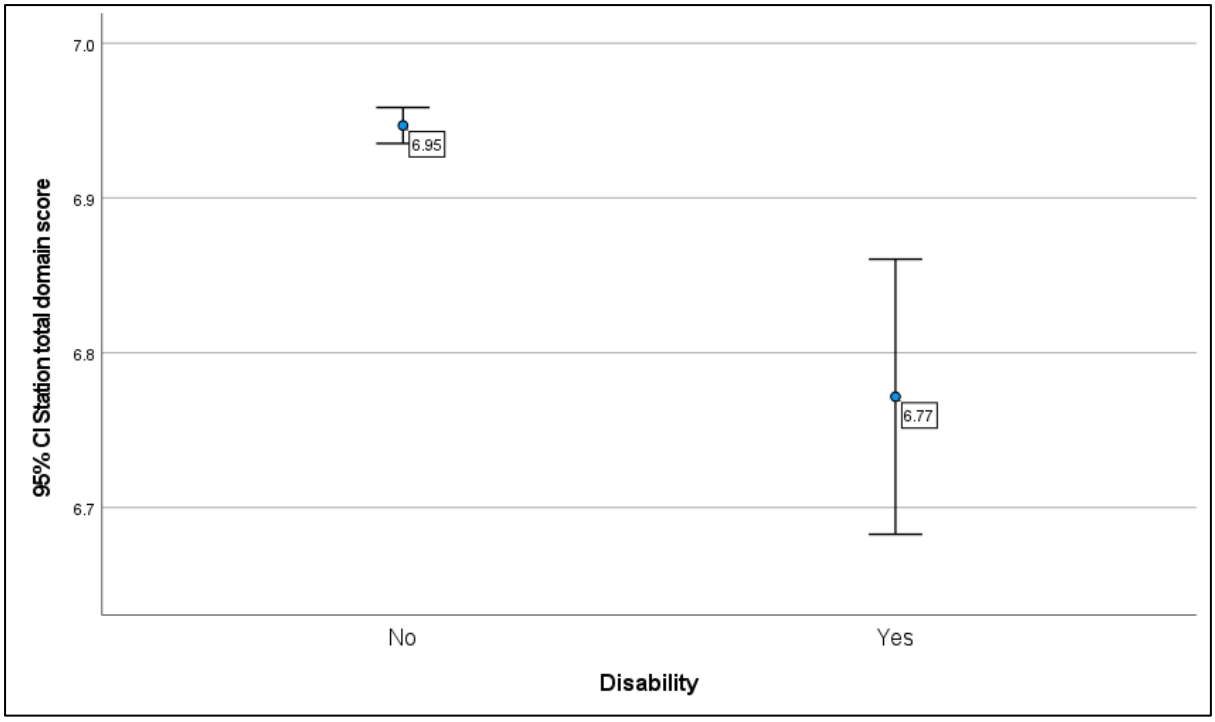

Figure 8: Error bar of total domain scores by examiner disability status

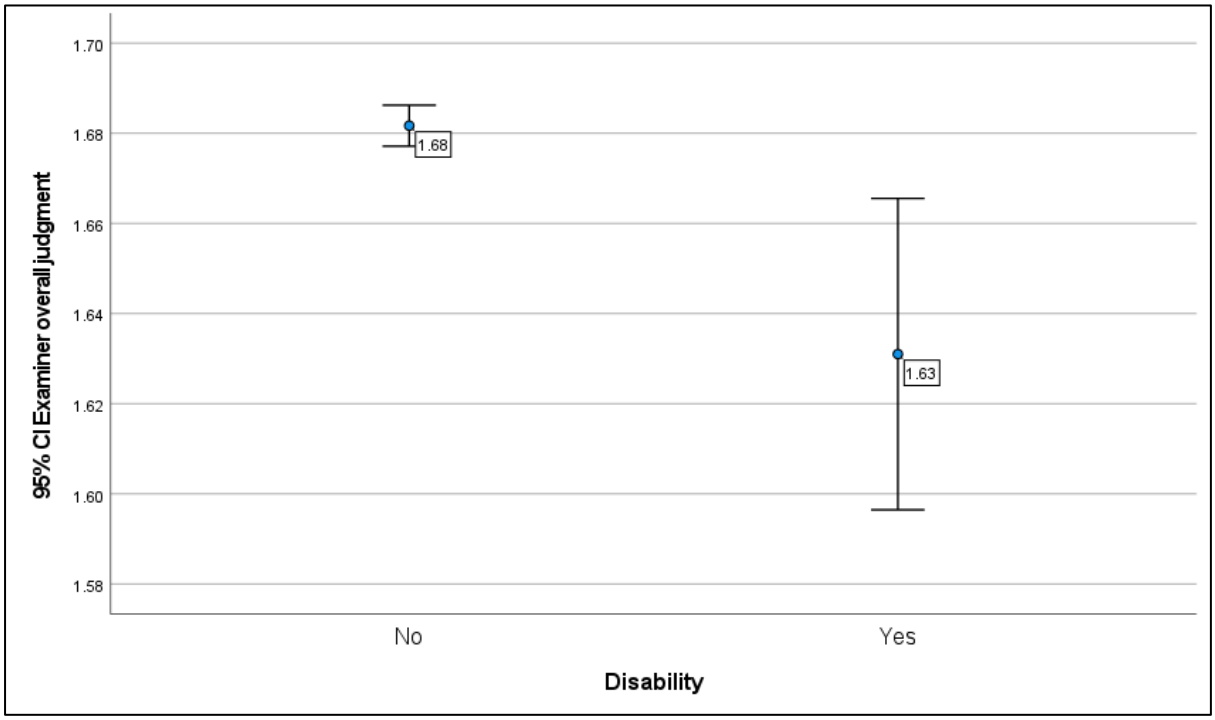

Figure 9: Error bar of global grades by examiner disability status

Examiner sexual orientation

There are some differences in scores and grades by *Examiner sexual orientation*, with gay male examiners tending to give lower scores and grades.

For scores see Figure 10 (F(2, 189859)=216.3, p<0.001, R squared=0.002), and for grade Figure 11 (F(2, 189859)=44.7, p<0.001, R squared=0.000).

However, the response category *Gay Man* forms a relatively small proportion of the data (2.9%).

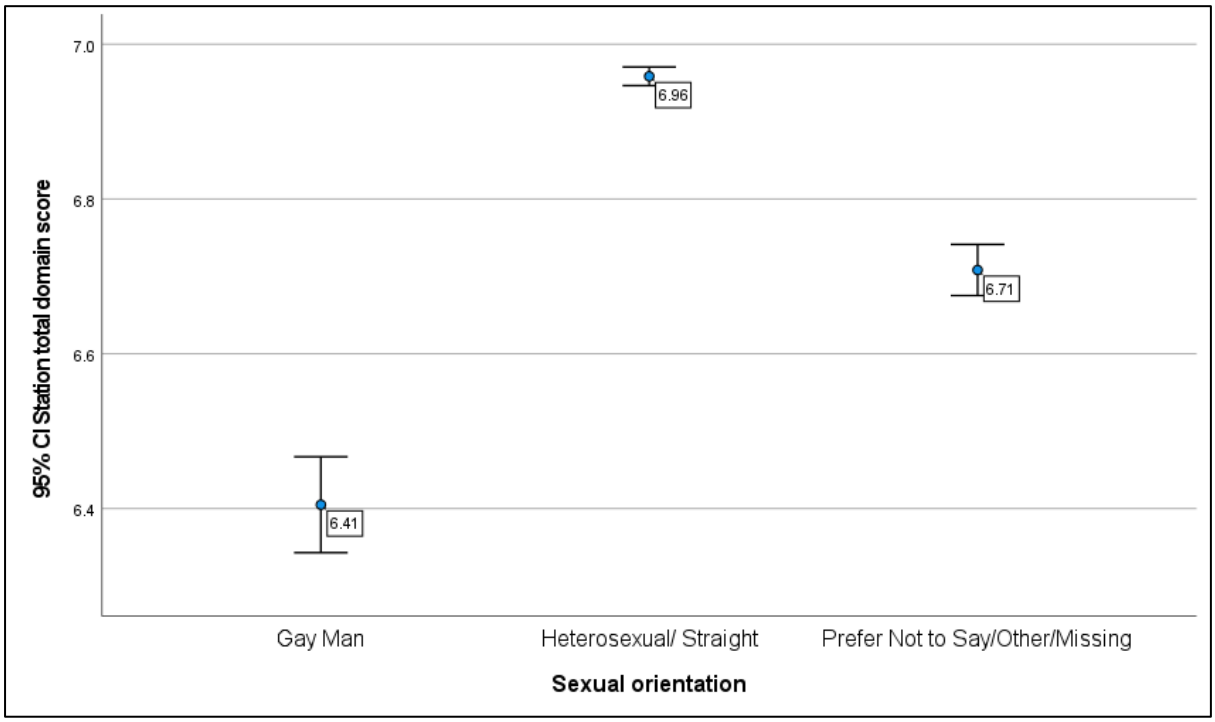

Figure 10: Error bar of total domain scores by examiner sexual orientation

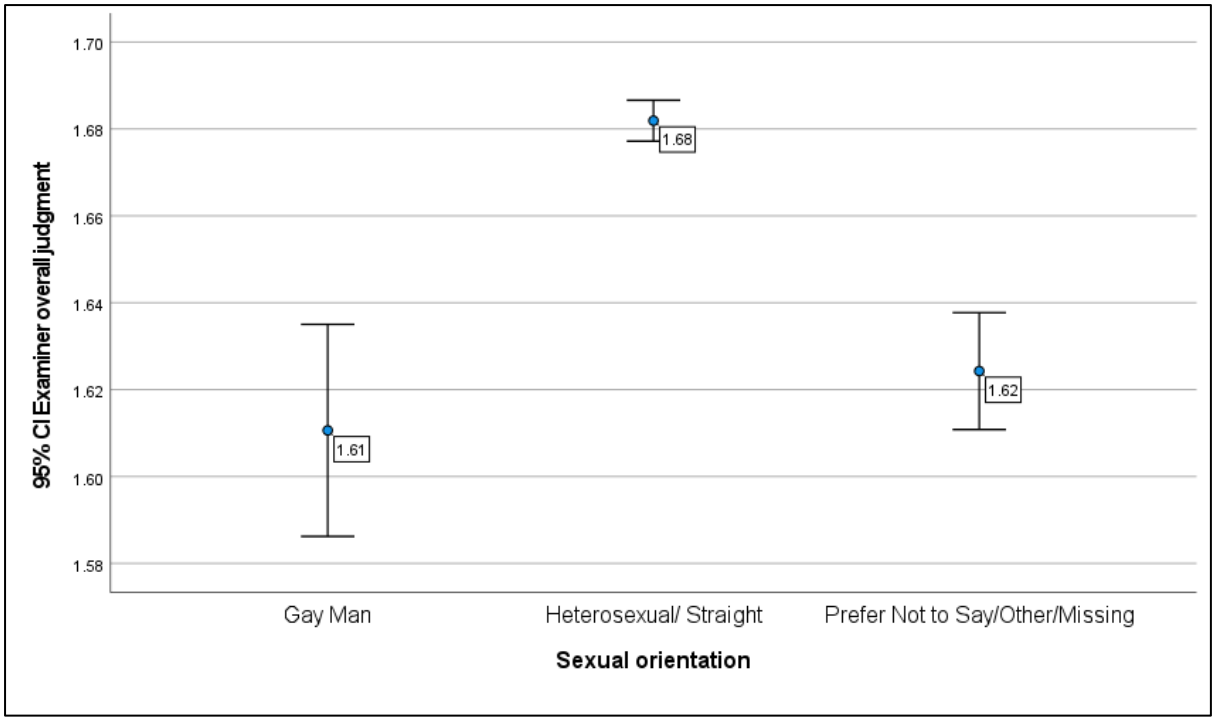

Figure 11: Error bar of global grades by examiner sexual orientation

Examiner religion

There are some differences in scores and grades by *Examiner religion,* but the overall effect is again quite small.

For scores, see Figure 12 ($F(5, 189856)=162.4$, $p<0.001$, R squared=0.004) and for global grades see Figure 13 ($F(5, 189856)=99.0$, $p<0.001$, R squared=0.003).

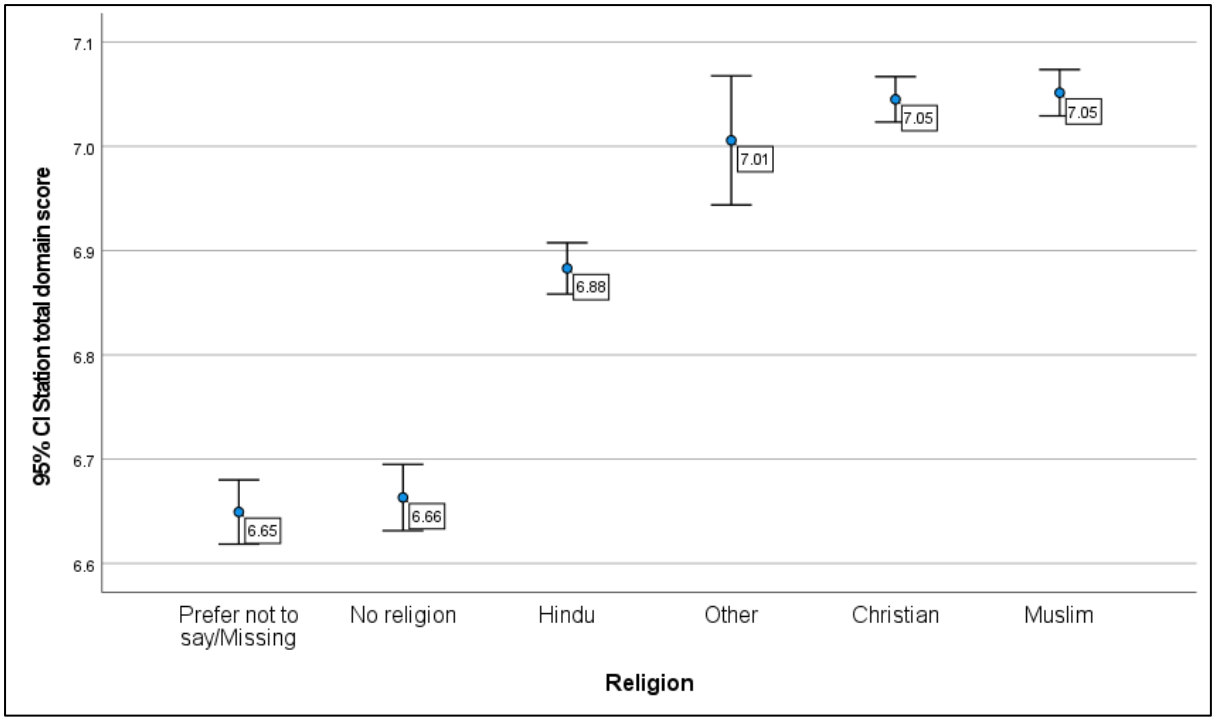

Figure 12: Error bar of total domain scores by examiner religion

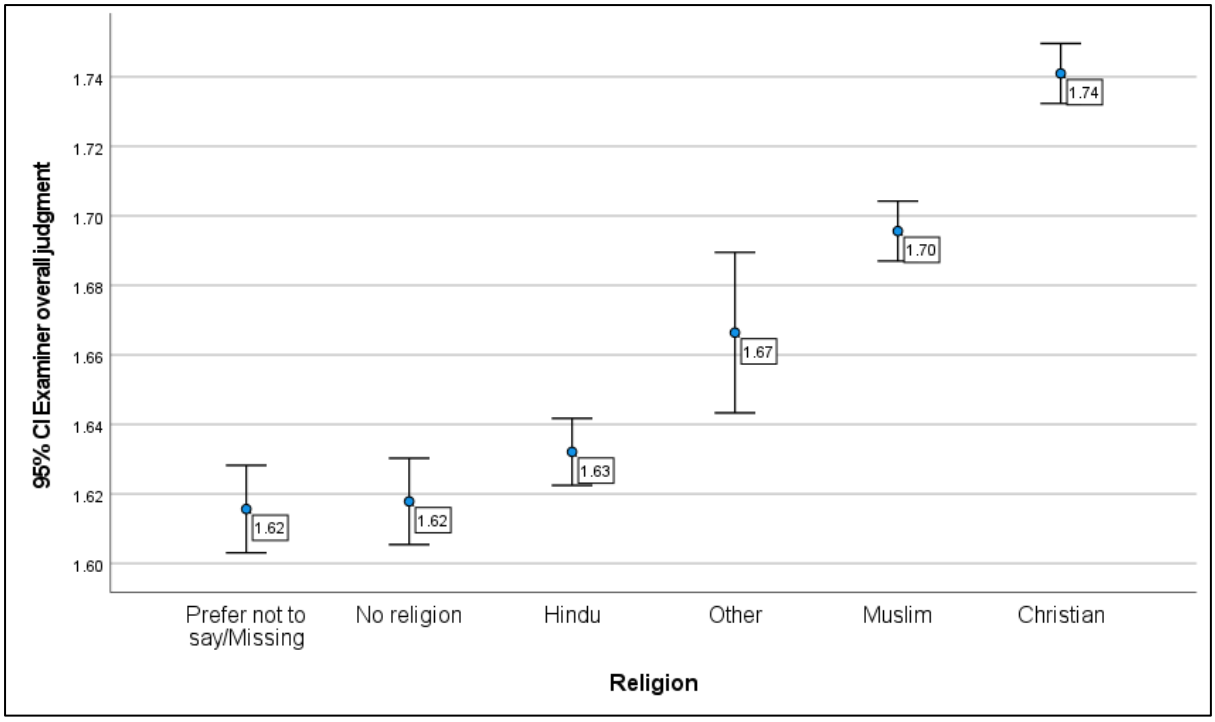

Figure 13: Error bar of global grades by examiner religion

Examiner date of first registration

Figure 14 (F(1, 189860)=560.1, p<0.001, R squared=0.003) and Figure 15 (F(1, 189860)=197.5, p<0.001, R squared=0.001) show that there is a slight upward trend in that those more recently registered tend to score more highly.

Note that this is a scale variable, so the appropriate figures are scatter plots.

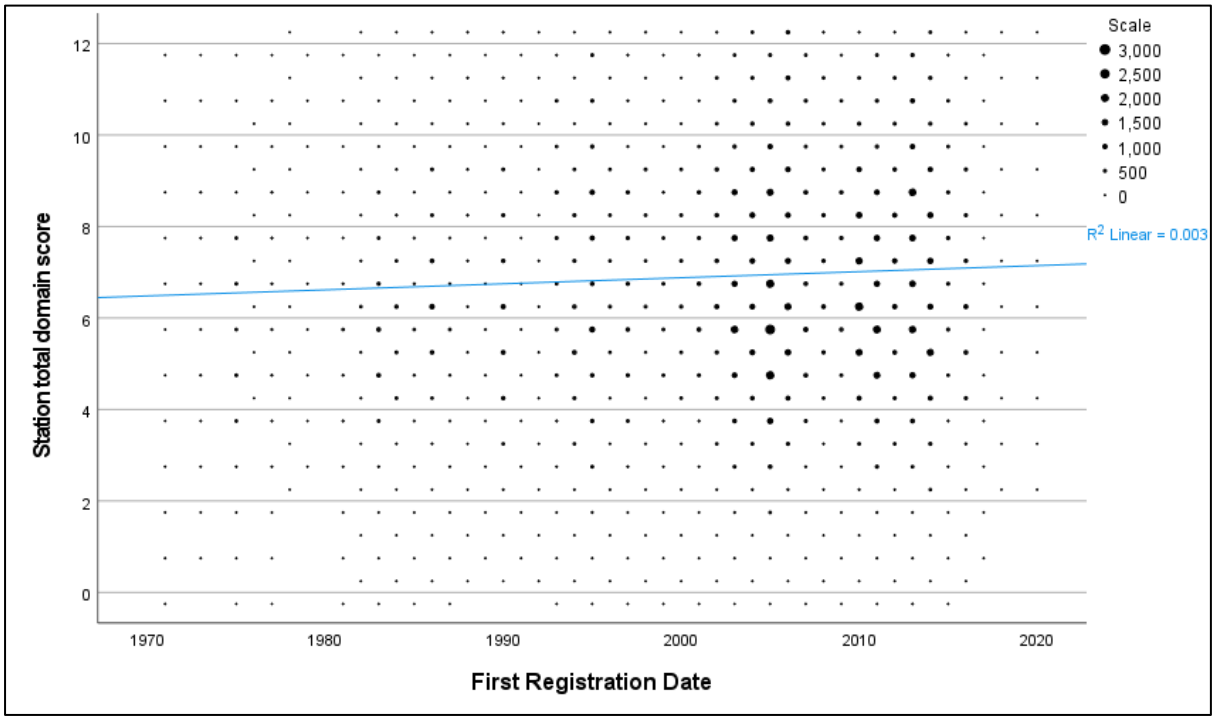

Figure 14: Scatter plot of total domain scores by examiner first registration date

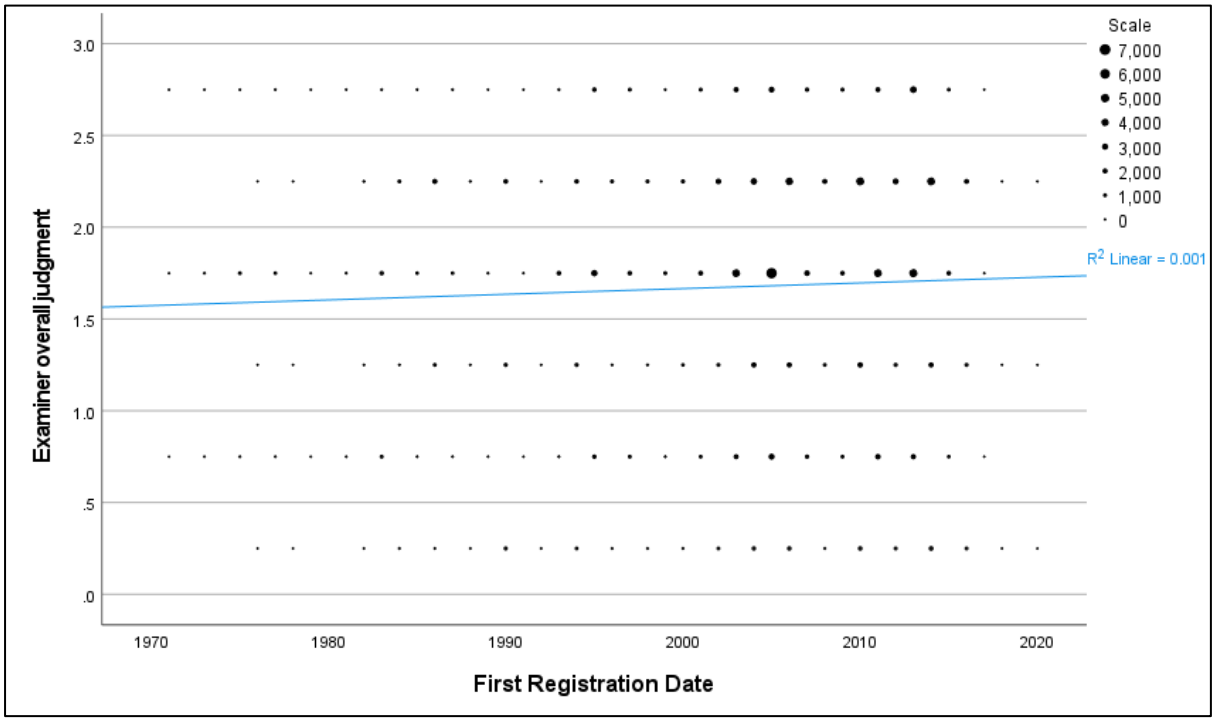

Figure 15: Scatter plot of global grades by examiner first registration date

Examiner status as GP

Figure 16 and Figure 17 show that there is almost no difference in mean scores by *Examiner status as a GP* (F(1, 189860)=0.9, p=0.34, R squared=0.000), F(1, 189860)=0.0, p=0.96, R squared=0.000).

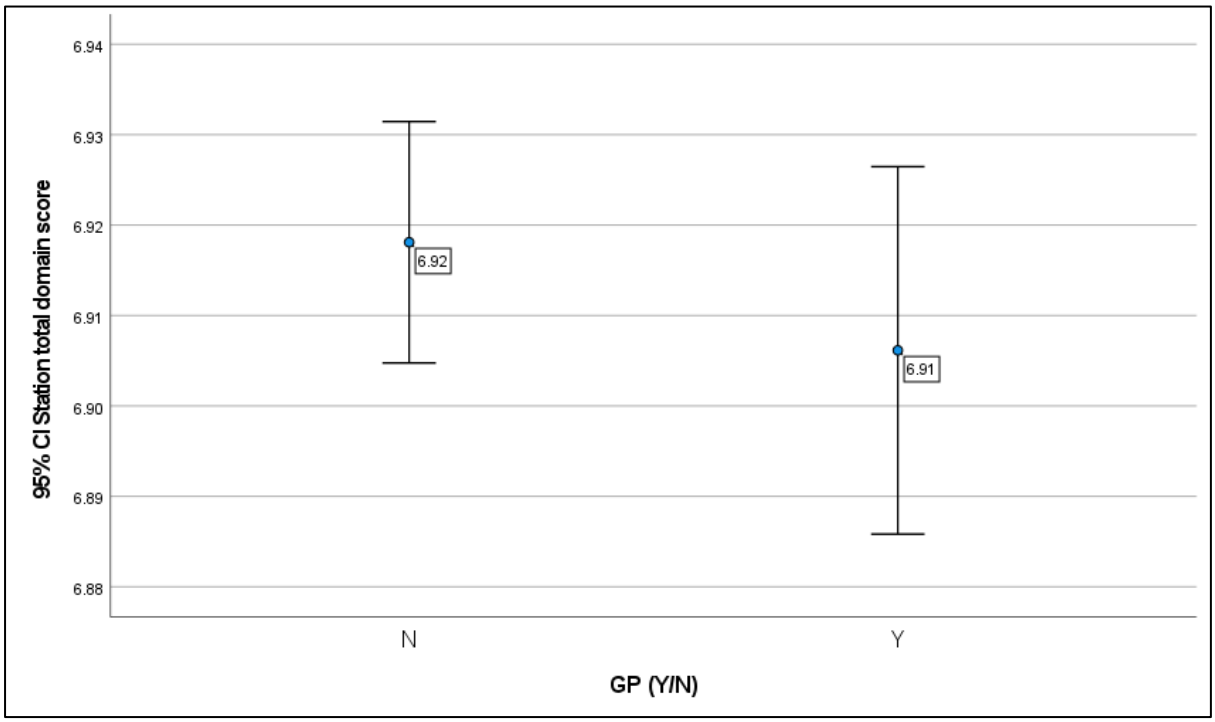

Figure 16: Error bar of total domain scores by examiner GP status

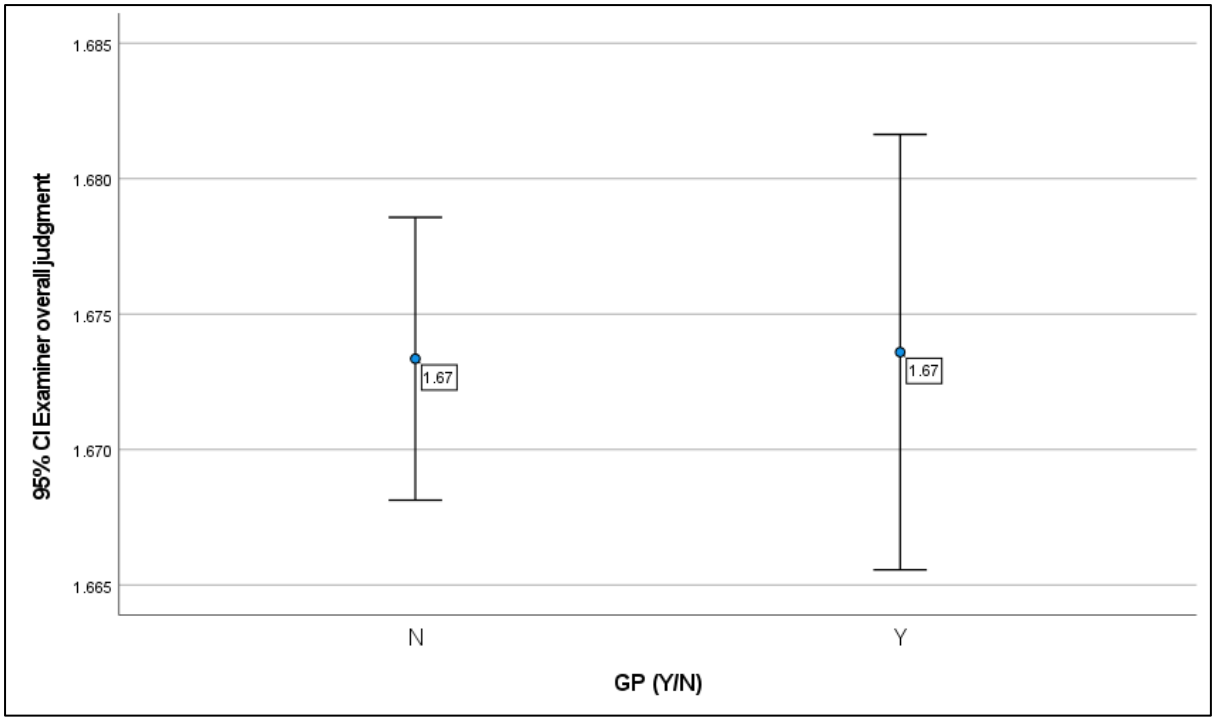

Figure 17: Error bar of global grades by examiner GP status

Examiner status as a specialist

Figure 18 indicates that there is a some difference in mean scores by *Examiner status as a specialist* – with non-specialists tending to score more highly (F(1,189860)=96.5, p<0.001, R squared=0.001).

Interestingly, the pattern is in the opposite directions, but less strong, for global grades – see Figure 19 (F(1,189860)=29.7, p<0.001, R squared=0.000).

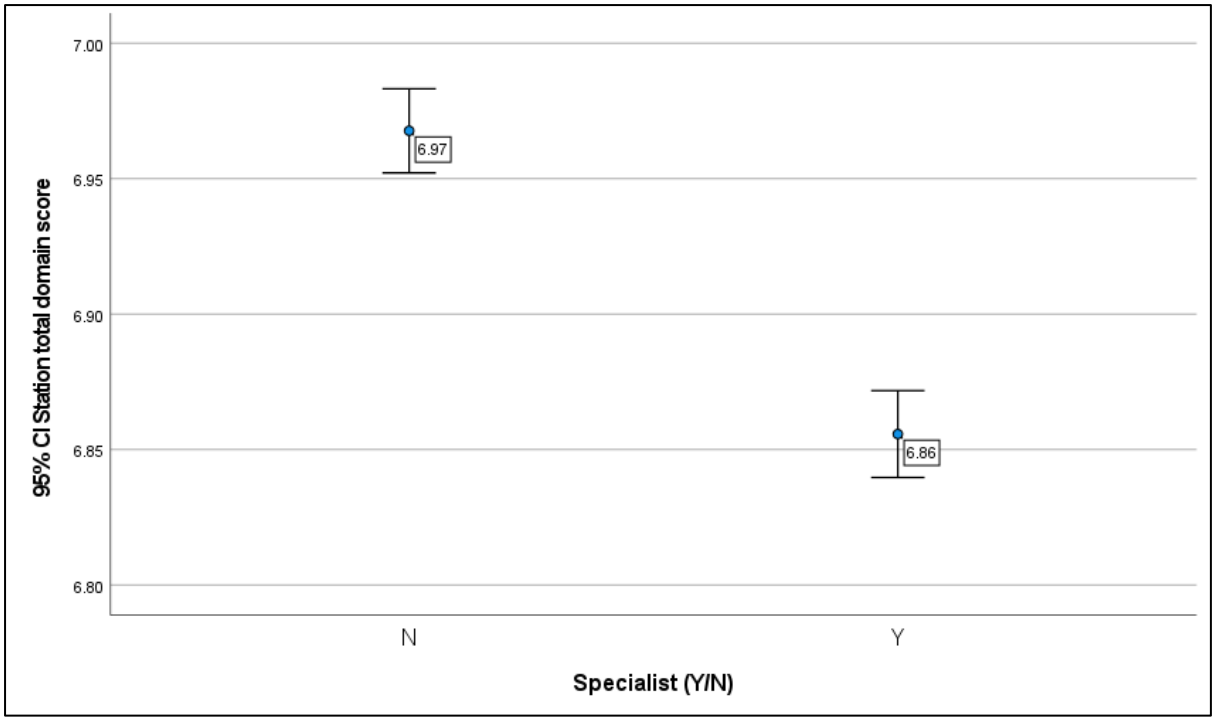

Figure 18: Error bar of total domain scores by examiner specialist status

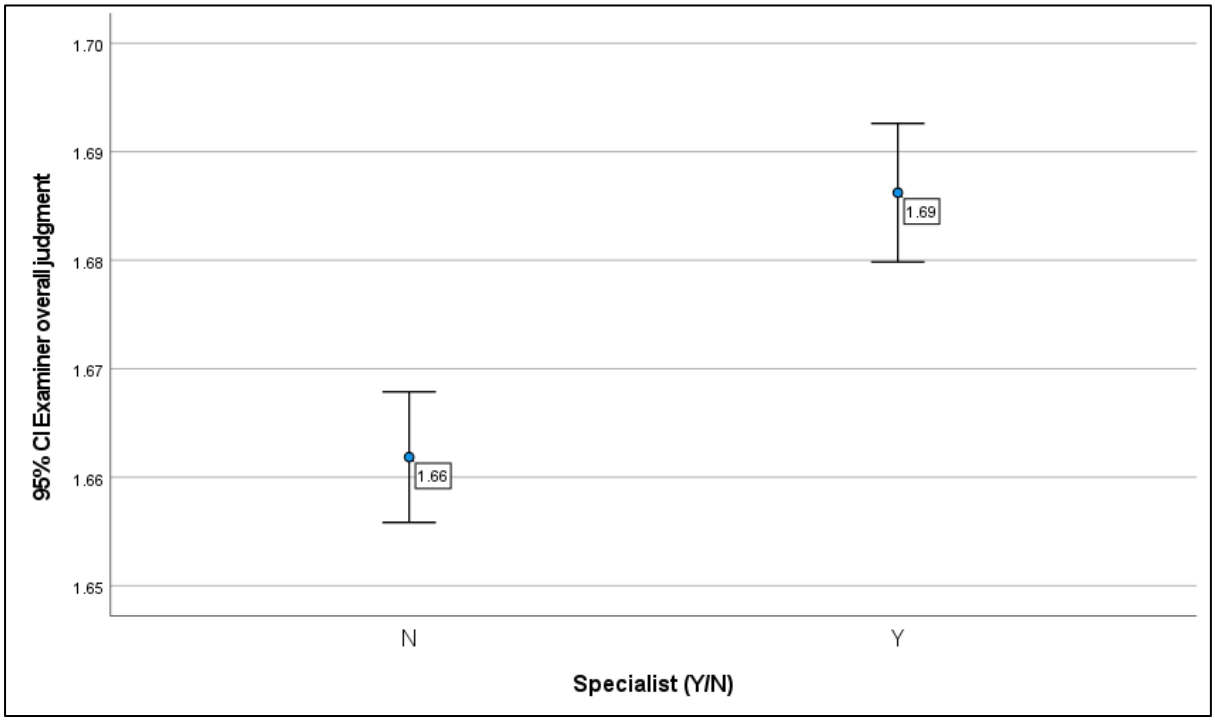

Figure 19: Error bar of global grades by examiner specialist status

Examiner country of Primary medical qualification

Figure 20 and Figure 21 show the mean score (and 95% confidence intervals) for scores and grades respectively within each PMQ country. Table 3 gives essentially the same data organised alphabetically by PMQ country.

The GLM statistics for these two analysis are as follows:
$F(45, 189816)=78.6$, $p<0.001$, R squared=0.018) for scores, and
$F(45, 189816)=64.3$, $p<0.001$, R squared=0.015) for grades

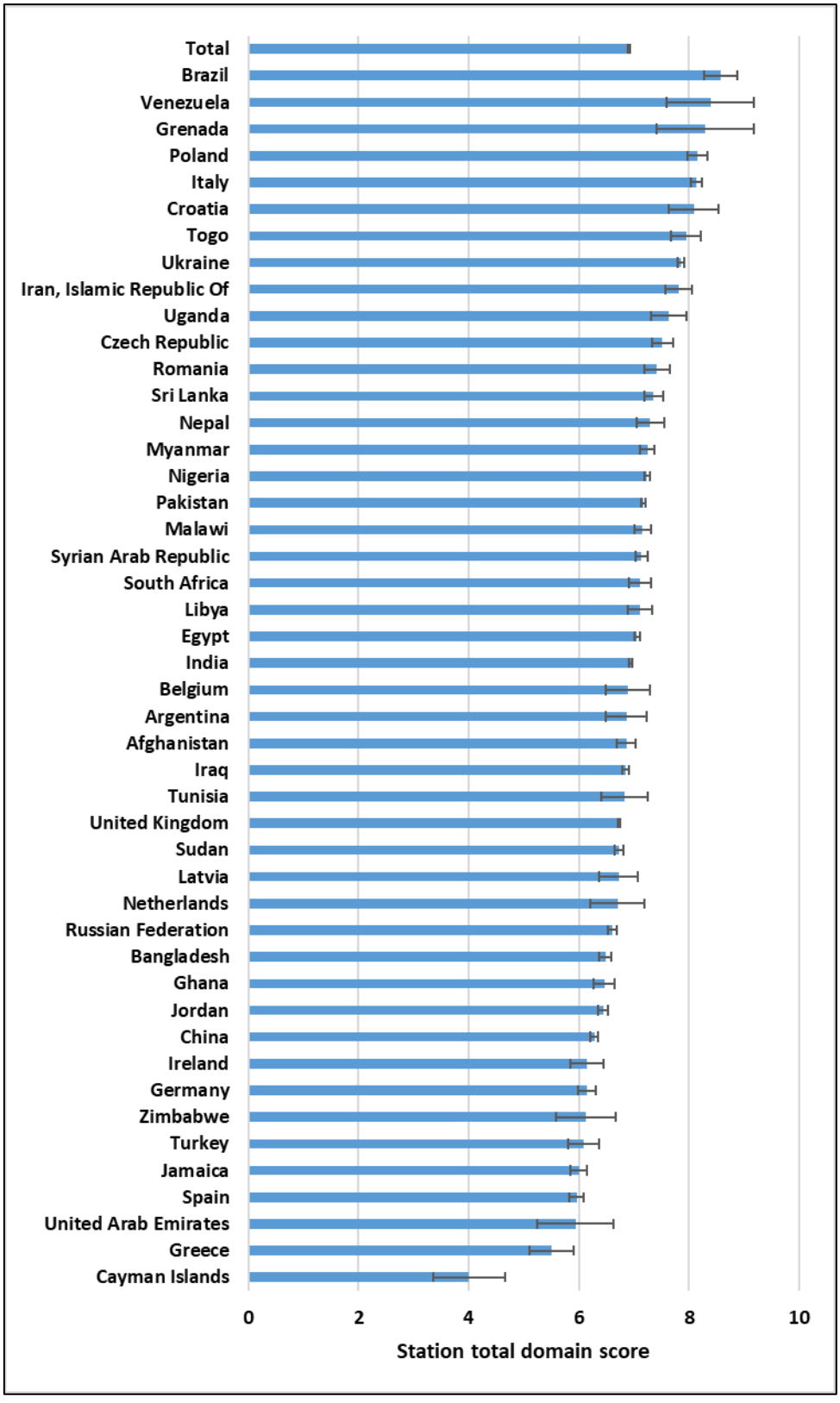

Figure 20: Error bar of total domain scores by Examiner PMQ country of origin

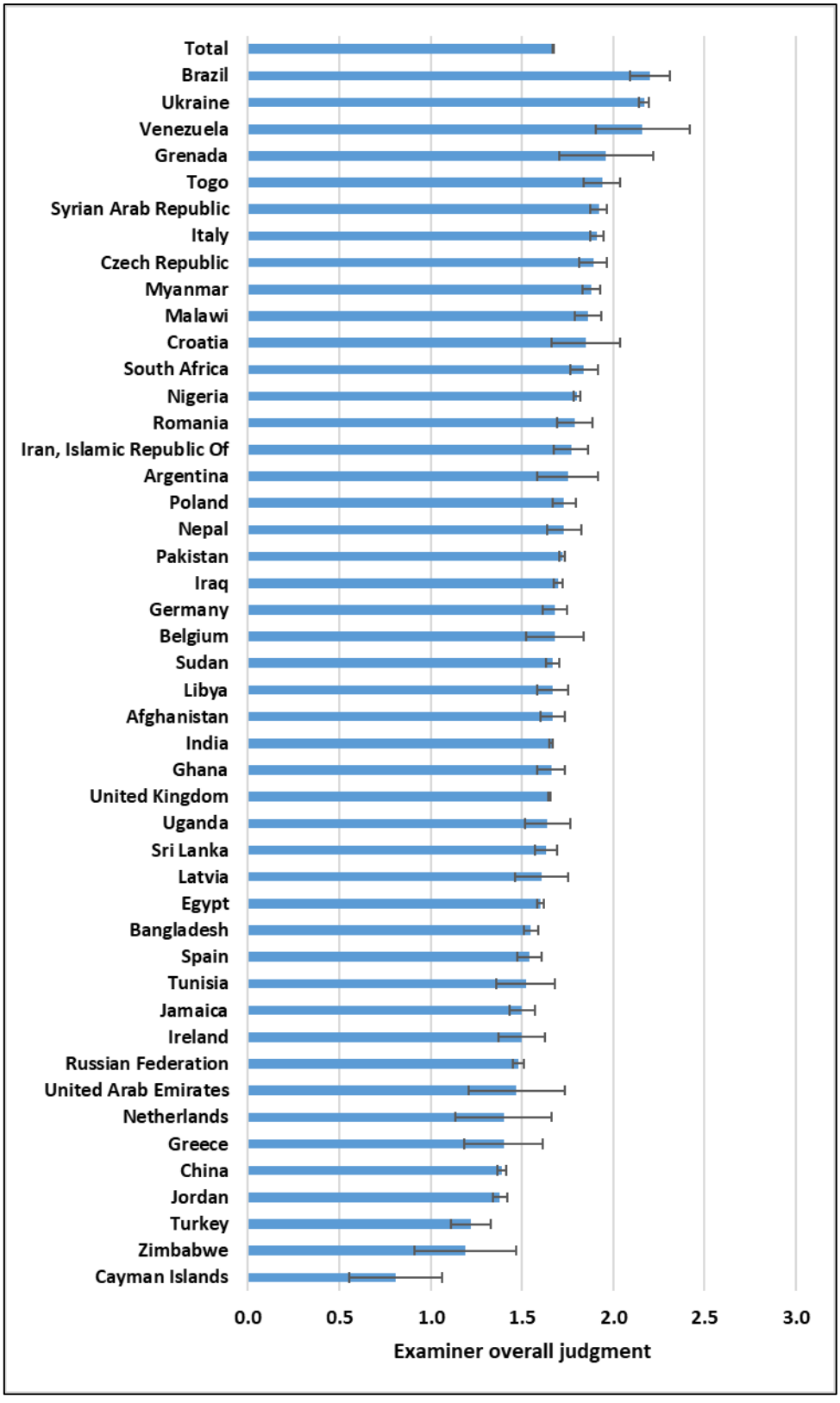

Figure 21: Error bar of global grades by Examiner PMQ country of origin

PMQ Country	Mean score	Mean grade	N
Afghanistan	6.86	1.67	511
Argentina	6.86	1.75	183
Bangladesh	6.48	1.55	2,304
Belgium	6.89	1.68	152
Brazil	8.58	2.2	301
Cayman Islands	4.00	0.81	32
China	6.28	1.39	4,652
Croatia	8.09	1.85	141
Czech Republic	7.52	1.89	605
Egypt	7.06	1.6	8,497
Germany	6.14	1.68	701
Ghana	6.46	1.66	522
Greece	5.50	1.4	62
Grenada	8.29	1.96	28
India	6.94	1.66	48,237
Iran, Islamic Republic Of	7.81	1.77	490
Iraq	6.85	1.7	6,310
Ireland	6.14	1.5	273
Italy	8.13	1.91	2,031
Jamaica	6.00	1.5	627
Jordan	6.44	1.38	2,522
Latvia	6.72	1.61	119
Libya	7.11	1.67	635
Malawi	7.16	1.86	294
Myanmar	7.25	1.88	1,174
Nepal	7.30	1.73	471
Netherlands	6.70	1.4	57
Nigeria	7.24	1.8	11,702
Pakistan	7.17	1.72	15,287
Poland	8.15	1.73	764
Romania	7.42	1.79	470
Russian Federation	6.61	1.48	3,583
South Africa	7.11	1.84	497
Spain	5.96	1.54	736
Sri Lanka	7.36	1.63	1,052
Sudan	6.73	1.67	2,503
Syrian Arab Republic	7.14	1.92	1,265
Togo	7.95	1.94	398
Total	6.91	1.67	189,862
Tunisia	6.83	1.52	180
Turkey	6.09	1.22	363
Uganda	7.63	1.64	277
Ukraine	7.85	2.17	4,221

United Arab Emirates	5.94	1.47	32
United Kingdom	6.73	1.65	64,539
Venezuela	8.39	2.16	31
Zimbabwe	6.13	1.19	31
Total	**6.91**	**1.67**	**189,862**

Table 3: Summary statistics for station scores and grades by Examiner PMQ country of origin

Station type

As might have been hypothesised, there are some differences in scores and grades by station type – with *Prescription* stations scoring/grading the lowest, whilst *Standard* stations have the highest mean values.

For scores, see Figure 22 ($F(4, 189857)=1432.9$, $p<0.001$, R squared=0.038) and for global grades Figure 23 ($F(4, 189857)=1876.0$, $p<0.001$, R squared=0.029).

These R-squared values suggest that station type is more important in terms of influencing scores and grades compared to all of the fixed effects investigated so far.

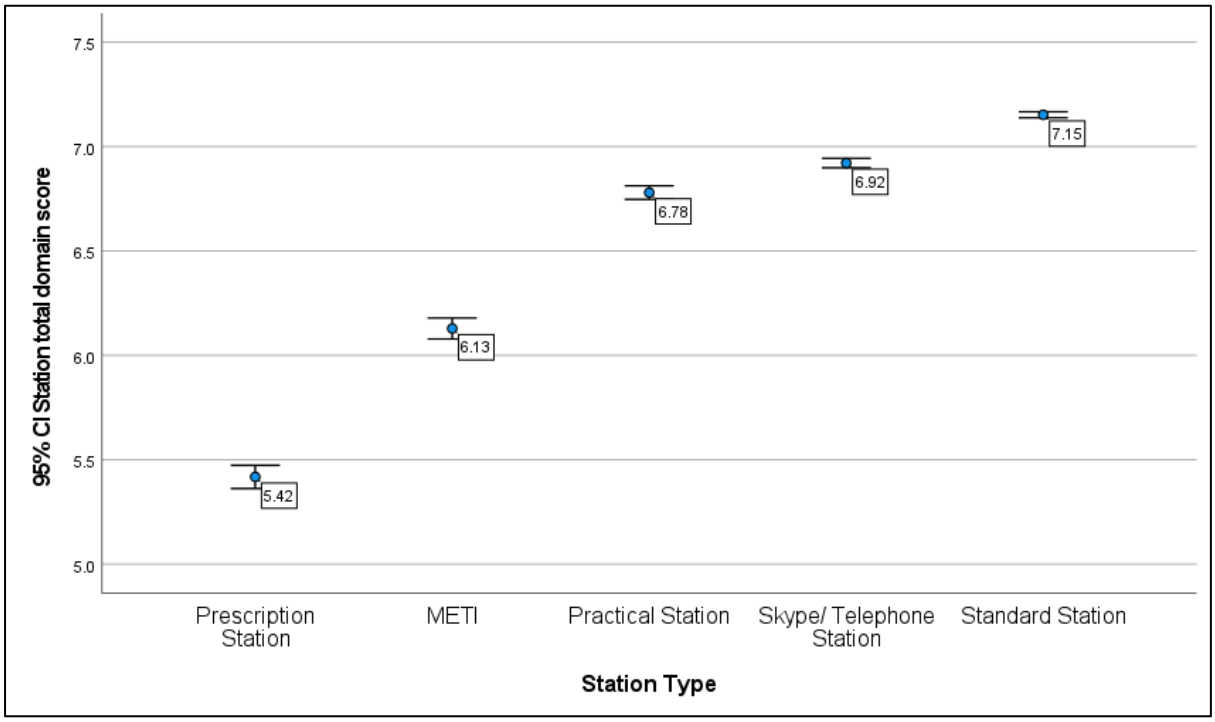

Figure 22: Error bar of total domain scores by station type

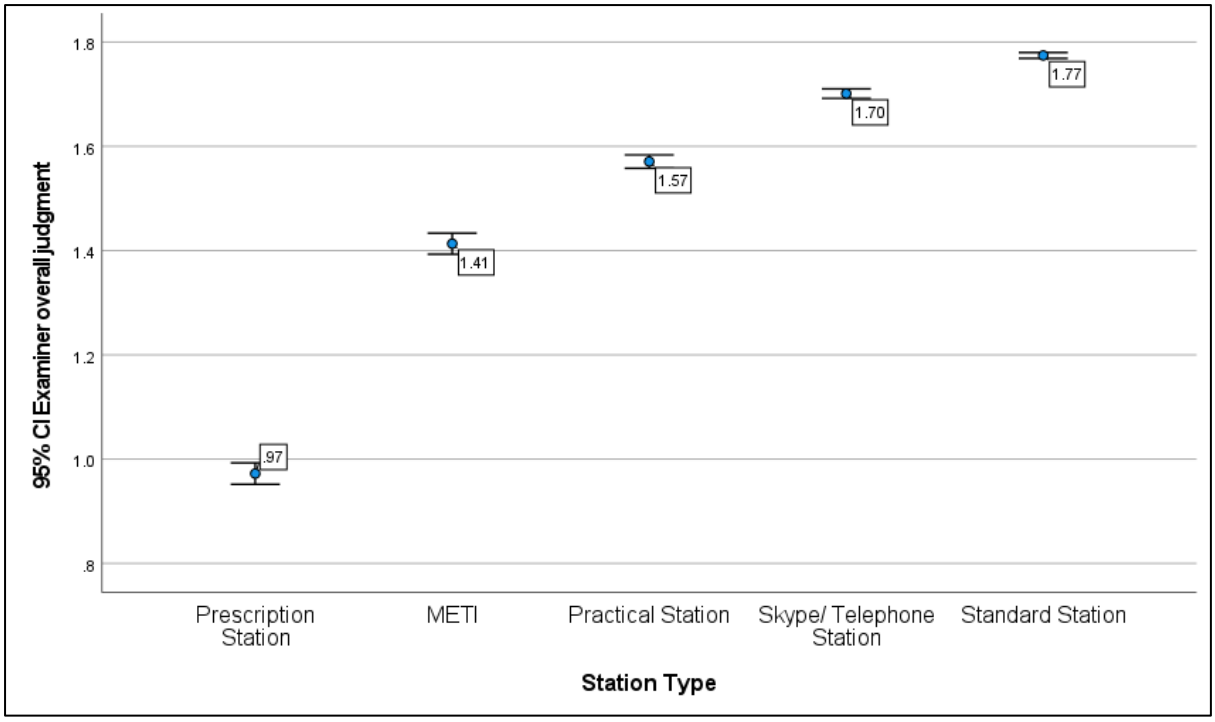

Figure 23: Error bar of global grades by station type

Summary of individual fixed effect influences on scores and grades

For convenience, we summarise in Table 4 the R squared values for the fixed effects analyses. These show how low the predictive power of most of these factors are for scores and grades. One can use the following informal guidelines to interpret the size of each of these effects: R-squared=0.01 is 'small', 0.06 is 'medium', and 0.14 is 'large'.[4]

Station type is by far the most predictive (R squared=0.038 and 0.029 for scores and grades respectively – so small to medium according to the usual guidelines). Other than for *Examiner PMQ country of origin (*also small to medium*)*, other effects are all small according to this guidance.

[4] See 4th row of table at https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize. However, these are somewhat arbitrary, and effect size guidance is best considered to be context specific.

Fixed effect	R squared	
	Score	Grade
Examiner sex	0.000	0.000
Examiner ethnicity	0.006	0.005
Examiner disability	0.000	0.000
Examiner sexual orientation	0.002	0.000
Examiner religion	0.004	0.003
Examiner date of first registration	0.003	0.001
Examiner status as GP	0.000	0.000
Examiner status as specialist	0.001	0.000
Examiner PMQ country of origin	0.018	0.015
Station type	0.038	0.029

Table 4: Percentages of variance in scores and grades for fixed effects

Individual influences on scores and grades – random effects

As described in the Methodology, it is appropriate to treat *Candidate*, *Examiner*, *Station*, *Exam* and *Examiner PMQ country of origin* as random effects – essentially each of these variables has many levels, and the actual levels (i.e. values) in the data can be thought of as a random sample from the theoretically infinite set of all possible levels.[5]

Table 5 shows the results of a set of null models predicting scores or grades using the single random effect as a predictor. As we might have expected, based on previous work on PLAB2 and other OSCE-type assessments (Yeates and Sebok-Syer, 2017; Homer, 2022; Homer, 2023b) we see that *Examiner* is the strongest predictor of total domain scores in stations followed by *Candidate, Examiner PMQ country of origin*, and *Station*.

Exam makes very little contribution to variation in scores – this makes sense as we would not expect that there would be much systematic variation in scores across separate exams based solely on this factor.

We should point out that the purpose of an assessment is usually to discriminate between candidates. Hence, *Candidate* variance in total domain scores is seen as 'good', whereas most or all other variance is 'bad' (i.e. is likely to be interpreted as error in the

[5] PMQ country is, arguably, different in this regard but methodologically it still makes sense to treat it as a random effect.

measurement/assessment). Whilst the fact that the *Examiner* effect is more strongly predictive of scores than, say, *Candidate*, might therefore seem a serious concern, it should be remembered that the analysis has all been carried out at the station level. At the exam level, however, individual examiner differences will tend to cancel out, at least to an extent, and the influence of *Candidate* will become stronger at the full exam level (Homer, 2022).

For grades, the pattern in Table 5 is somewhat similar to that of scores, but with typically less variance explained for each random effect. We will revisit this finding in the Discussion and conclusion.

Random effect	Percentage of variance attributable to random effect in null model	
	Score	Grade
Candidate	10.3	9.3
Examiner	17.3	10.5
Station	9.5	9.5
Exam	1.5	1.1
Examiner PMQ country of origin (see also Figure 20 and Figure 21)	9.0	4.9

Table 5: Percentages of variance in scores and grades for random effects in separate models

The results in Table 5 suggest that is important in the more complex modelling to correctly estimate the effect of each of these random effects in a combined model (see next section). It is possible that some effects will be ameliorated in such a model.

If we now compare the findings in Table 5 with the strength of the fixed effects in Table 4, we see that percentage of variance in PLAB2 outcomes explained by the random factors are typically much larger than for the fixed effects. However, the modelling and treatment of variables across the two types of analysis are different so direct comparison is not entirely straightforward. That said, this results so far as strongly suggestive that the random effects are more important influences on PLAB2 outcomes than are the fixed effects.

We move on now to the more complex, multivariate modelling where both fixed and random effects are included in the models.

Multiple influences on scores and grades

We use a combined (multivariate) model to investigate independent influences on PLAB2 scores (and then grades) of the examiner and station characteristics, as well as other factors, as detailed separately in the previous sub-sections. This will give us estimates of the separate effects of each characteristic on the scores (and grades) having controlled for the others.

We start with a model only including the random effects, and then employ all effects, both fixed and random.

Random effects only model

Table 6 gives the components of variance in scores and grades when we run a model that only includes all random effects (*Candidate*, *Examiner*, *Station*, *Exam* and *Examiner PMQ country of origin)*.

We see that there is a lot of unexplained variance - 66% and 73% for scores and grades respectively. This tells us that whilst factors like *Candidate*, *Examiner* and *Station* are important in influencing scores/grades (all of the order of 10% of variance explained), there could be other, possibly unmeasured, factors that are also important.

From Table 6 it is also clear that *Exam* and *Examiner PMQ country of origin* are not important in influencing station level PLAB2 outcomes above and beyond the other factors included.

Random effect	Percentage of variance attributable to random effect in random effects only model	
	Score	Grade
Candidate	10.3	9.3
Examiner	15.1	8.5
Station	7.8	8.5
Exam	0.5	0.4
Examiner PMQ country of origin	0.3	0.1
Residual	66.0	73.1
Total	100.0	100.0

Table 6: Percentages of variance in scores and grades in random effects only model

There are mainly small shifts in the components of variance when comparing models with a single random effect (Table 5) to that in a combined model (Table 6). The exception to this is *Examiner PMQ country of origin* – with much larger components in a single model (e.g. for scores, 9.0% vs 0.3% in the combined model). The much small components for this factor in the combined model strongly suggest that differences across PMQ countries of origin are actually mostly due to differences in other (random) factors, most likely *Examiner* itself - given that *Examiner* and *PMQ* are confounded in the sense that for each examiner the PMQ country is fixed (in other words, examiners are nested in PMQ countries). Whilst these arguments are a little methodologically obtuse, the clear conclusion to be drawn here is that *Examiner PMQ country of origin* plays almost no role in influencing scoring/grading above and beyond the other random effects included in the modelling.

A final observation concerning the findings of this random effects only model is that the pattern of there being larger estimates for scores compared to grades, as witnessed in Table 5, is also generally the case in Table 6.

The full (multivariate) model with all fixed and random effects as predictors

On adding all the fixed effects (*Examiner sex*, *Examiner ethnicity*, …) to produce a full model, the percentages of variation due to the random factors change little (compare the values in Table 7 to those already seen in Table 6 – they are very similar).

Random effect	Percentage of residual variance attributable to random effect in full model	
	Score	Grade
Candidate	10.6	9.7
Examiner	15.6	8.8
Station	5.6	5.7
Exam	0.5	0.4
Examiner PMQ country of origin	0.2	0.0
Residual	67.5	75.4
Total	100.0	100.0

Table 7: Percentages of variance in scores and grades in full model

This tells us that having accounted for a range of additional factors, there is little change in the balance of importance of the random factors in driving scoring/grading. This is probably to be expected given that the individual influences of each of the fixed effects on scores/grades has already been shown to be generally quite small (Table 4).

There is also a lot of unexplained variance – which most likely due to limitations in the modelling approach. For example, we are necessarily assuming a single candidate ability, when we know that many candidates vary in their performance across stations or task. Hence, at the station-level we should not be surprised that there is a lot of variation in scoring/grading that is not accounted for by the factors we have in the model.

The independent estimates for the influence of each level of each fixed effect on the scoring/grading in PLAB2 in the multivariate model are shown in Table 8. Only variables where at least one of the estimates on scores or grades is statistically significant at the 5% level have been included (and shaded), but for the full modelling results see the Appendix.

For interpretation, each estimate gives the average model-predicted change in scores (or grades) when comparing between the category listed and the reference category. For example, in the first data row of Table 8 we see that the score estimate for *Hindu* versus *Christian* examiners is -0.35, equivalent on the 12-point scale to -2.92%. In other words, the model suggests that having accounted for other factors, *Hindu* examiners score around 3% lower on average compared to *Christian* examiners.

There are a total of 27 fixed effect estimates in the full model (as shown in Appendix), but Table 8 only includes five of these – so most effects are not statistically significant in influencing PLAB2 scores (or grades) – this includes for example, *Examiner ethnicity*.

As we might have expected from the simpler analyses (Figure 22, Figure 23), *Station type* is a relatively strong influence on PLAB2 scores and grades. Of the other fixed effects, only one level of *Examiner religion* (*Hindu*) is statistically significant in influencing scores when comparing with the reference group for this variable (*Christian*).

Fixed effects		Score		Grade		Interpretation
		Estimate (%)	p-value	Estimate (%)	p-value	
Examiner religion – Hindu (reference group *Christian*)		-0.35 (-2.92)	0.023	-0.07 (-2.29)	0.113	*Hindu* examiners are more hawkish on scoring compared to the *Christian* examiners – by just under 3% on average. For grades, the pattern is similar but does not reach the 5% cut-off for statistical significance.
Station type (reference group *Standard*)	**METI**	-0.86 (-7.19)	<0.0001	-0.31 (-10.27)	<0.0001	All station types score lower compared to the *Standard* stations – by between around 2% (*Skype/telephone*) to 16% (*Prescription*). There is a similar pattern by grade with larger effects – equivalent figures are 2% and 28% respectively).
	Practical	-0.41 (-3.43)	<0.0001	-0.21 (-6.92)	<0.0001	
	Prescription	-1.88 (-15.66)	<0.0001	-0.83 (-27.66)	<0.0001	
	Skype/telephone	-0.20 (-1.68)	0.002	-0.06 (-2.08)	0.012	

Table 8: Fixed effects (significant at p=0.05 level) for scores/grades in full model

Discussion and conclusion

This work set out to investigate a range of potential influences on PLAB2 station-level outcomes – both total domain scores, and examiner global judgments. The main finding is that most factors, those we have referred to as fixed effects, do not play a significant role in influencing station-level outcomes. In particular, in the combined modelling almost no examiner characteristics (e.g. ethnicity or sex) are found to have any relationship with PLAB2 outcomes at the station level. These 'null' (i.e. negative) findings are important, as we might have hoped in advance that this would indeed be the case – if they had shown to be important this would have been seen as contributing error to the exam, thereby threatening the validity of the overall pass/fail outcomes (Cook et al., 2015). A rigorous analysis has identified that only a single level of one factor (the religion of the examiner) is significant in influencing outcomes (and in fact, only in the case of total domain scores, not grades).

The only fixed effect factor that has an important impact on PLAB2 outcomes across all its different categories is the type of station being assessed – and results are as might have been hypothesised in advance: stations classified as *Standard* score and grade more highly on average compared to other station types, with stations classified as *Prescription* being the lowest scoring.

That said, there are factors in the exam that do influence PLAB2 station-level outcomes, the most important of which are the examiner themselves, the candidate and the station being administered - these are referred to throughout this report as random effects because these factors in the data can each be thought of as being a sample from a wider potential population. The analysis shows, then, that examiners vary in their stringency, candidates in their ability, and stations in their difficulty – when all other factors in the modelling have been accounted for (i.e. controlled for to an extent). These are already well-established and expected findings given the evidence base that exists for PLAB2 (Homer, 2022; Homer, 2023b) and in other OSCE-type settings (Yeates and Sebok-Syer, 2017).

However, when station-level effects are aggregated up to the exam level, the apparent concern that examiners have a strong relative impact on scores is largely ameliorated. In this report we have not explicitly measured the extent of this – but it has been shown in other work where the components of variance by examiners relative to candidates was very similar (Homer, 2022).

That stations are also important in scoring/grading underlines the need to make the conjunctive minimum station hurdle at the exam level more defensible. If a particular exam is made up of a difficult set of stations, this work suggests that the fixed hurdle is equivalent to a higher standard than it would be for an easier set of stations – which is clearly unfair. Moving away from a fixed standard to one that takes account of the mix of station difficulties in an exam will help improve the quality and fairness of overall pass/fail PLAB2 outcomes (Homer, 2023a). Pilot work in this area is currently ongoing (Hankins et al., 2024).

There are few additional policy-oriented suggestions that derive from this work, but there are one or two findings perhaps worthy of additional comment. First, there is a pattern across most analyses that variation by fixed or random effects is typically a little larger for total domain scores in comparison to global grades (see for example Table 4, Table 6 and Table 7). The reason why this might be the case is not possible to say from the data to hand, but could be related to the more holistic nature of the single overall global judgment - but this needs more research to develop our understanding.

Another intriguing finding is that in the simple analysis of scores/grades by *Examiner status as a specialist.* The work indicates that specialists are on average more hawkish than non-specialists for total domain scores, but the difference is the other way around for global grades (see Figure 18 and Figure 19). That these differences are weakened in the more complex modelling (Table 9 and Table 10) suggests that it is other differences between specialists and non-specialists captured in the other factors included in the combined model that are driving such differences – but this would require additional analyses to confirm in detail.

There are some important limitations to this work. We have already mentioned the model assumptions of a single stringency, ability, and difficulty for examiners, candidates and stations respectively. This is something that could be investigated in further research with a subset of the data – i.e. for examiners/stations with relatively high (i.e. more frequent) occurrence in the data.

Finally, we have treated global grades as scale in nature throughout the analysis, when in reality these judgments are formally ordinal. However, given the large sample size in the analysis, and the fact that it is common to treat these grades in this way, not least under the borderline regression method of standard setting commonly used for these types of assessments (McKinley and Norcini, 2014), it seems reasonable to assume that a more nuanced approach would not reveal much of additional import.

References

Amrhein, V., Greenland, S. and McShane, B. 2019. Scientists rise up against statistical significance. *Nature*. **567**(7748), p.305.

Bates, D., Mächler, M., Bolker, B. and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. **67**(1), pp.1–48.

Cook, D.A., Brydges, R., Ginsburg, S. and Hatala, R. 2015. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education*. **49**(6), pp.560–575.

Field, A. 2013. *Discovering Statistics Using IBM SPSS Statistics* 4th edition. Sage Publications.

Hankins, R., Homer, M. and Caballero, J. 2024. Assessing a dynamic method of setting the conjunctive standard for a major licensing osce.

Homer, M. 2022. Pass/fail decisions and standards: the impact of differential examiner stringency on OSCE outcomes. *Advances in Health Sciences Education*.

Homer, M. 2023a. Setting defensible minimum-stations-passed standards in OSCE-type assessments. *Medical Teacher*., pp.1–7.

Homer, M. 2023b. Towards a more nuanced conceptualisation of differential examiner stringency in OSCEs. *Advances in Health Sciences Education*.

Hutchison, D. and Schagen, I. 2008. Concorde and discord: the art of multilevel modelling. *International Journal of Research & Method in Education*. **31**(1), p.11.

IBM Corp 2021. IBM SPSS Statistics for Windows, Version 28.0.

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.

Wasserstein, R.L., Schirm, A.L. and Lazar, N.A. 2019. Moving to a World Beyond "p < 0.05". *The American Statistician*. **73**(sup1), pp.1–19.

Yeates, P. and Sebok-Syer, S.S. 2017. Hawks, Doves and Rasch decisions: Understanding the influence of different cycles of an OSCE on students' scores using Many Facet Rasch Modeling. *Medical Teacher*. **39**(1), pp.92–99.

Appendix

For modelling scores and grades the following apply to the results:

- **Green** highlighting for absolute t-value less than 2 (usually associated with non-statistical significance).
- **Red** highlighting for p-values[6] less than 0.05, the usual cut-off value for statistical significance.

[6] p-values in mixed models are controversial but with large sample sizes are seen as not too problematic – see for example: https://www.r-bloggers.com/2014/02/three-ways-to-get-parameter-specific-p-values-from-lmer/

- Number of observations: 173823, groups: candidates, 12749; Examiners, 778; Stations, 532; Exams, 290; Examiner PMQ country of origin, 46
- Reference groups for categorical variables:
  - Examiner sex – female
  - Examiner ethnicity – Missing
  - Examiner disability – No
  - Examiner religion – Christian
  - Examiner status as GP – No
  - Examiner status as specialist – No
  - Station type – Standard

Full model fixed effects for scores

	Estimate	Estimate (%)	Std. Error	t value	p-value
(Intercept)	-3.38	-28.16	9.94	-0.34	0.377
Examiner sex – male	0.03	0.27	0.08	0.41	0.367
Examiner ethnicity - Asian / Asian British - Chinese	-0.40	-3.37	0.42	-0.97	0.250
Examiner ethnicity - Asian or Asian British - Bangladeshi	0.12	0.98	0.43	0.28	0.384
Examiner ethnicity - Asian or Asian British - Indian	0.10	0.86	0.38	0.27	0.384
Examiner ethnicity - Asian or Asian British - Other	-0.06	-0.47	0.41	-0.14	0.395
Examiner ethnicity - Asian or Asian British - Pakistani	0.02	0.18	0.39	0.05	0.398
Examiner ethnicity - Black or Black British - African	0.02	0.14	0.40	0.04	0.399
Examiner ethnicity - Other	-0.01	-0.08	0.40	-0.02	0.399
Examiner ethnicity - Other ethnic group – Arab	0.09	0.79	0.41	0.23	0.388
Examiner ethnicity - White - British, English, Northern Irish, Scottish	-0.20	-1.64	0.38	-0.52	0.348
Examiner ethnicity - White - Other	0.10	0.86	0.40	0.26	0.386
Examiner disability – Yes	-0.26	-2.17	0.23	-1.15	0.207
Examiner sexual orientation - Gay Man	0.10	0.85	0.26	0.40	0.369
Examiner sexual orientation - Prefer Not to Say/Other/Missing	0.09	0.71	0.16	0.52	0.348
Examiner religion – Hindu	-0.35	-2.92	0.15	-2.40	0.023
Examiner religion – Muslim	-0.06	-0.50	0.14	-0.42	0.365
Examiner religion – No religion	-0.15	-1.21	0.14	-1.06	0.227
Examiner religion – Other	-0.13	-1.09	0.23	-0.58	0.337
Examiner religion - Prefer not to say/Missing	-0.32	-2.71	0.19	-1.74	0.088
Examiner – First Registration Date	0.01	0.05	0.00	1.10	0.217
Examiner status as GP – Yes	-0.15	-1.24	0.10	-1.46	0.137
Examiner status as specialist – Yes	-0.17	-1.45	0.10	-1.69	0.095
Station type - METI	-0.86	-7.19	0.16	-5.28	0.000
Station type - Practical	-0.41	-3.43	0.10	-4.00	0.000
Station type - Prescription	-1.88	-15.66	0.13	-14.11	0.000
Station type - Skype/ Telephone	-0.20	-1.68	0.06	-3.34	0.002

Table 9: Fixed effects in full model for scores

Full model fixed effects for grades

	Estimate	Estimate (%)	Std. Error	t value	p-value
(Intercept)	-0.43	-14.31	2.929	-0.147	0.395
Examiner sex – male	0.02	0.64	0.024	0.809	0.288
Examiner ethnicity - Asian / Asian British - Chinese	-0.16	-5.34	0.123	-1.307	0.170
Examiner ethnicity - Asian or Asian British - Bangladeshi	0.04	1.27	0.124	0.306	0.381
Examiner ethnicity - Asian or Asian British - Indian	-0.04	-1.37	0.110	-0.376	0.372
Examiner ethnicity - Asian or Asian British - Other	-0.07	-2.32	0.119	-0.585	0.336
Examiner ethnicity - Asian or Asian British - Pakistani	0.00	-0.10	0.113	-0.026	0.399
Examiner ethnicity - Black or Black British - African	-0.01	-0.17	0.115	-0.044	0.399
Examiner ethnicity - Other	-0.02	-0.73	0.116	-0.189	0.392
Examiner ethnicity - Other ethnic group – Arab	-0.05	-1.64	0.118	-0.415	0.366
Examiner ethnicity - White - British, English, Northern Irish, Scottish	-0.05	-1.76	0.110	-0.478	0.356
Examiner ethnicity - White - Other	0.01	0.34	0.116	0.089	0.397
Examiner disability – Yes	-0.10	-3.30	0.067	-1.472	0.135
Examiner sexual orientation - Gay Man	0.07	2.19	0.076	0.863	0.275
Examiner sexual orientation - Prefer Not to Say/Other/Missing	0.02	0.57	0.049	0.349	0.375
Examiner religion – Hindu	-0.07	-2.29	0.043	-1.588	0.113
Examiner religion – Muslim	-0.03	-1.16	0.042	-0.827	0.283
Examiner religion – No religion	-0.04	-1.39	0.041	-1.029	0.235
Examiner religion – Other	-0.07	-2.20	0.066	-0.994	0.243
Examiner religion - Prefer not to say/Missing	-0.10	-3.34	0.055	-1.817	0.077
Examiner – First Registration Date	0.00	0.04	0.001	0.786	0.293
Examiner status as GP – Yes	-0.02	-0.76	0.030	-0.768	0.297
Examiner status as specialist – Yes	-0.03	-1.00	0.030	-0.993	0.244
Station type - METI	-0.31	-10.27	0.064	-4.835	0.000
Station type - Practical	-0.21	-6.92	0.040	-5.162	0.000
Station type - Prescription	-0.83	-27.66	0.052	-15.947	0.000
Station type - Skype/ Telephone	-0.06	-2.08	0.024	-2.635	0.012

Table 10: Fixed effects in full model for grades