PLAB 1 and 2 Annual Report 2019 - April 2019

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

Contents

List of figures

List of tables

Executive summary

Using PLAB1 and PLAB2 data from 2019, with four and 174 test administrations respectively (Table 2), this report has the following key findings.

Some mismatch between Angoff judgements and item performance in PLAB1

- At the item level, candidate performance for the borderline group in PLAB1 does not generally match that expected according to the *a priori* Angoff judgments. Typically, borderline performance is much more spread-out than expected (see for example, Figure 8). However, at the test level these differences tend to even out allaying concern about the overall standard set. It is worth considering whether the Angoff process, and judgments made, might need reviewing, perhaps with additional encouragement for judges to use a wider range of expected item facilities for the borderline group. Recent changes to the make-up of the panel, including trainees, might also have a beneficial effect in this regard.

Little evidence of deterioration in candidate performance for later items in PLAB1

- There is little evidence that item performance for the candidate group as a whole deteriorates significantly for items positioned later in the PLAB1 test (Figure 9). The same is true for the borderline group (Figure 10). Items very near the end of the test do sometimes have higher rates of non-response for the borderline group, but these rates are still quite low - of the order of 2% non-response (Figure 15).

A moderately strong correlation between candidate performance on PLAB1 and PLAB2

- There is a moderately strong correlation between PLAB1 and PLAB2 scores – ($r=0.53$, $n=8,704$, $p<0.001$) – having adjusted for missing data and measurement error (Figure 21 and associated text). The magnitude of the correlation is probably of the order of what one might expect – with performance in PLAB1 clearly predicting, to an extent, PLAB2, but the lack of perfect correlation also indicating that the two PLAB tests are not measuring the same construct(s).

No important deterioration in assessment quality with the advent of two circuits in PLAB2

- The advent of two circuits in PLAB2 from August 2019 has led to a decline in average reliability as measured by Cronbach's alpha across administrations from 0.75 to 0.72 ($p=0.007$, Cohen's $d=0.45$; Table 5). The most likely cause of this change is that there are now two assessors in each station, rather than being due to any other aspect of the new PLAB2 arrangements. However, given that a standard error of measurement (SEM) is added to the pass mark when calculating the pass mark, any decrease in reliability increases the SEM accordingly, and keeps the false positive rate constant. Other metrics also show a minor difference – for example, mean R-squared also has declined slightly (from 0.76 to 0.74, $p<0.001$, Cohen's $d=0.21$). As is the case for alpha, this is what we might expect on moving from a single examiner in a station to one with two sets of examiner scores/grades combined to set the standard under borderline regression.

- Comparison of various metrics across parallel circuits in PLAB2 indicate that, for most stations across administrations, there are no serious issues with differences in patterns of scoring across these (Table 6). This is more evidence that the move to two parallel circuits has not unduly affected the quality of the assessment outcomes of PLAB2.

No evidence of difference in examiner behaviour comparing morning to afternoon in PLAB2

- A comparison of scores for morning versus afternoon in PLAB2 does not indicate any substantial evidence of change in examiner behaviour later in the day (see, for example, Figure 27). In other words, there is no evidence in this analysis of examiner fatigue being a threat to the validity of the assessment or its outcomes.

Introduction

This report uses PLAB 1 and 2 data from 2019 to investigate the following research questions (RQs):

RQ1. What is the extent of the disparity between Angoff and actual facility for items from PLAB1?

RQ2. Does candidate performance in PLAB1 decay over the course of each exam, and is this worse for borderline candidates?

RQ3. What is the relationship between PLAB 1 and PLAB2 scores?

RQ4. What has the impact of the shift to two circuits been on key PLAB2 outcomes (i.e. measures of assessment quality, and candidate outcomes)?

RQ5. To what extent does time of day in PLAB2 (i.e. morning/afternoon) affect candidate outcomes?

Metrics glossary

There is a range of assessment/psychometric terminology used in this report – much of it quite specific to the borderline regression method (BRM) of standard setting used in PLAB2. To aid the reader, we detail a few key technical terms used as follows:

Level	Metric	Description
Test	**Cronbach's alpha**	Alpha is a commonly used measure of internal consistency reliability of a test. One interpretation is that alpha gives a measure of the correlation of the test with a hypothetical 'perfect' test – so values near 1 are better (i.e. indicate a more reliable test).
	Standard error of measurement (SEM)	This is a measure of how much error there is in a candidates score – and it depends, in part, on the reliability of the test with higher reliability corresponding to lower SEM and vice versa. In many assessment contexts, the SEM is of the order of 2-3% and in PLAB 1 and PLAB2 it is added to the cut-score to minimise the false positives in the test (that is borderline candidates passing because of error in their favour).
Station	**ANOVA R-squared**	This is a way of measuring the difference in domain scores in the same station but across parallel circuits. Under the assumption of candidates randomised to circuits, we can think of this as a proxy for the effect of assessors on these scores - so low values near 0 are good, and high values (i.e. over, say, 0.5) indicate, for example, the adverse effect of hawks/doves on domain scores.
	R-squared	Under BRM, R-squared is the correlation squared between the domain scores and the global grades in a station. This indicates the amount of shared variance between the two station level scores and typically we would hope for values (say) greater than 0.5 (Pell et al., 2010) in a well-functioning station.
	Slope	Under BRM, this is the slope of the regression line – and we don't want lines too flat or too steep. It is a measure of the average difference in domain scores between successive global grades.
	Intercept	Under BRM, this is the intercept where the regression line cuts the y-axis – so it is the projected average mark for the fail grade. We wouldn't want this negative, or too high.
Station or item	**Facility**	In PLAB1 this refers to the proportion of candidates that get the item correct (this is the standard definition in the wider literature). In PLAB2 this refers to the pass rate for the station.

Table 1: Glossary of key assessment/psychometric terms

Methodology

All PLAB1 and PLAB2 assessment data from the calendar year 2019 was analysed. Table 2 shows the sample sizes in the data:

Exam	No. of sittings	No. of candidates
PLAB1	4	11,114 (9,950 unique)
PLAB2	174	7,188 (5,877 unique)

Table 2: Number of sittings and candidates used in analysis

The data includes candidate scores/grades and pass/fail outcomes, test, station and item level cut-scores, and various additional metrics related to the assessments. A small number of suppressed items and stations were removed from the analysis as appropriate for certain analyses.

In terms of methods of analysis, as simple a methods as possible are used – often purely descriptive and graphical representations. On occasion the independent sample t-test is used to compare two groups, but with a focus on effect sizes rather than p-values (Wasserstein and Lazar, 2016). More specific individual methodological approaches are detailed at the appropriate points in the report.

Findings

RQ1 – Angoff versus item performance in PLAB1

For this RQ, we investigate the relationship between a priori Angoff judgements of item difficulty for the borderline candidate, and actual item performance. We begin with a brief comparison across the four administrations of the Angoff distributions.

Angoff distributions in each test

These are very similar across the four administrations – which suggests the tests are consistent in terms of patterns of expected item difficulties - see summary statistics and boxplots that follow:

Administration	Angoff total[1]	SEM	Cronbach alpha
March	115.6	5.59	0.87
June	114.3	5.70	0.88
September	115.8	5.23	0.89
November	114.1	5.34	0.90

Table 3: Angoff and other test level measures

[1] Includes a handful of supressed items – 3 in June and 3 in November

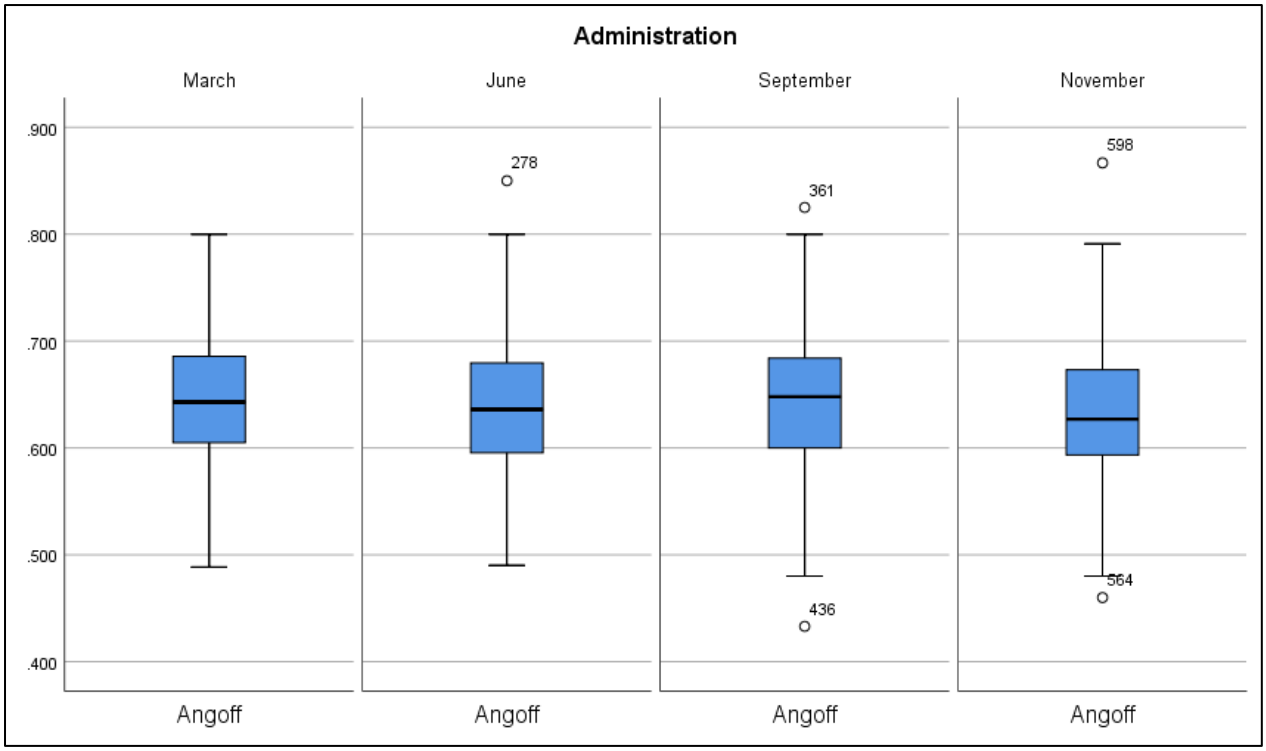

Figure 1: Box plots for Angoff judgements across the four 2019 administrations of PLAB1

Relationship between Angoff ratings and item performance

For the rest of the analysis for this RQ, we focus mainly here on March 2019 data – the substantive findings for other administrations are very similar.

We define the borderline group in PLAB1 as those within five marks of passing score (i.e. passing score ±1 SEM). In the March data this identifies 20.8% of the 4,286 candidates as borderline. We can then calculate the relationship between Angoff scores, item facility for all candidates[2], and item facility for this borderline group.

First we show histograms of these three item-level measures – with the key difference between them being that the Angoff judgments have a much narrower spread. Also, the difference in spread, and even central location (i.e. average) between the two facility distributions is not particularly great. This latter finding is because the full group of candidates in each sitting contains a wide range of performance – including large proportions of failures. So in this PLAB1 data, the full cohort is typically not that different in average performance than the borderline group.

[2] This group includes those failing, which is a substantial proportion of the cohort: 26, 38, 47 and 45% respectively over March to November 2019.

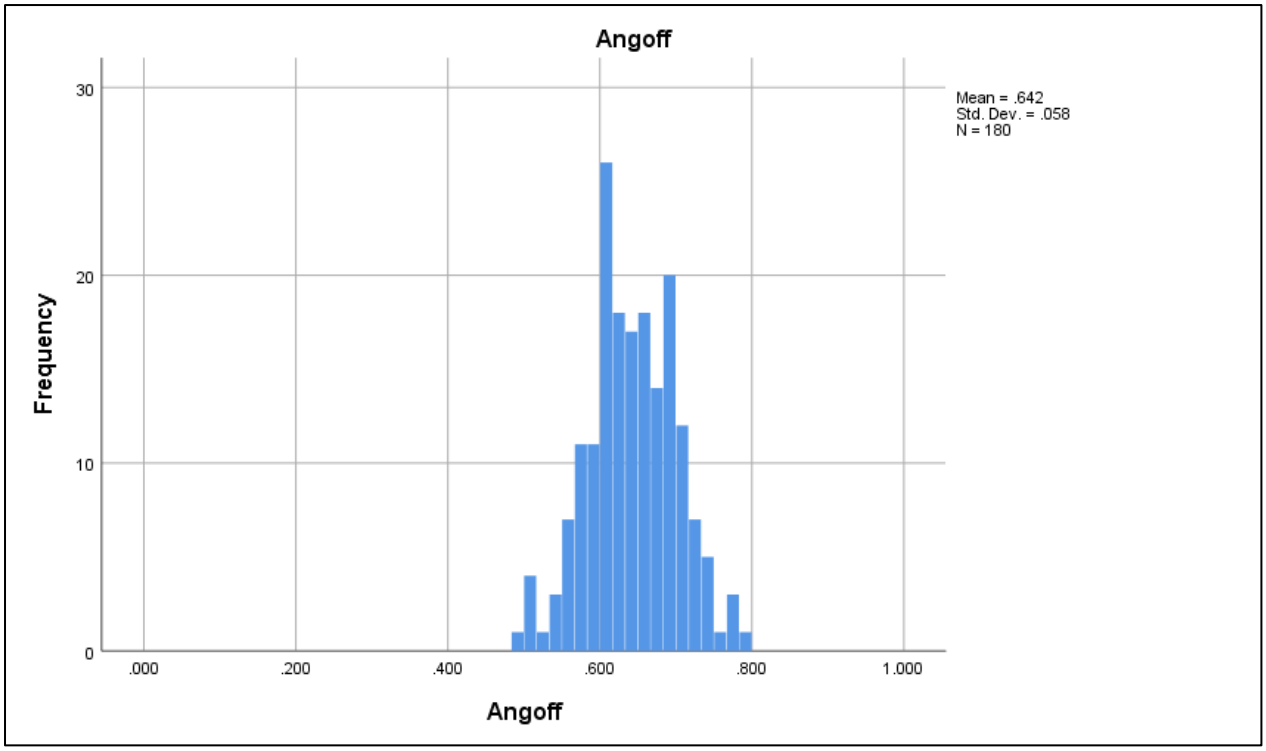

Figure 2: Distribution of Angoff judgments

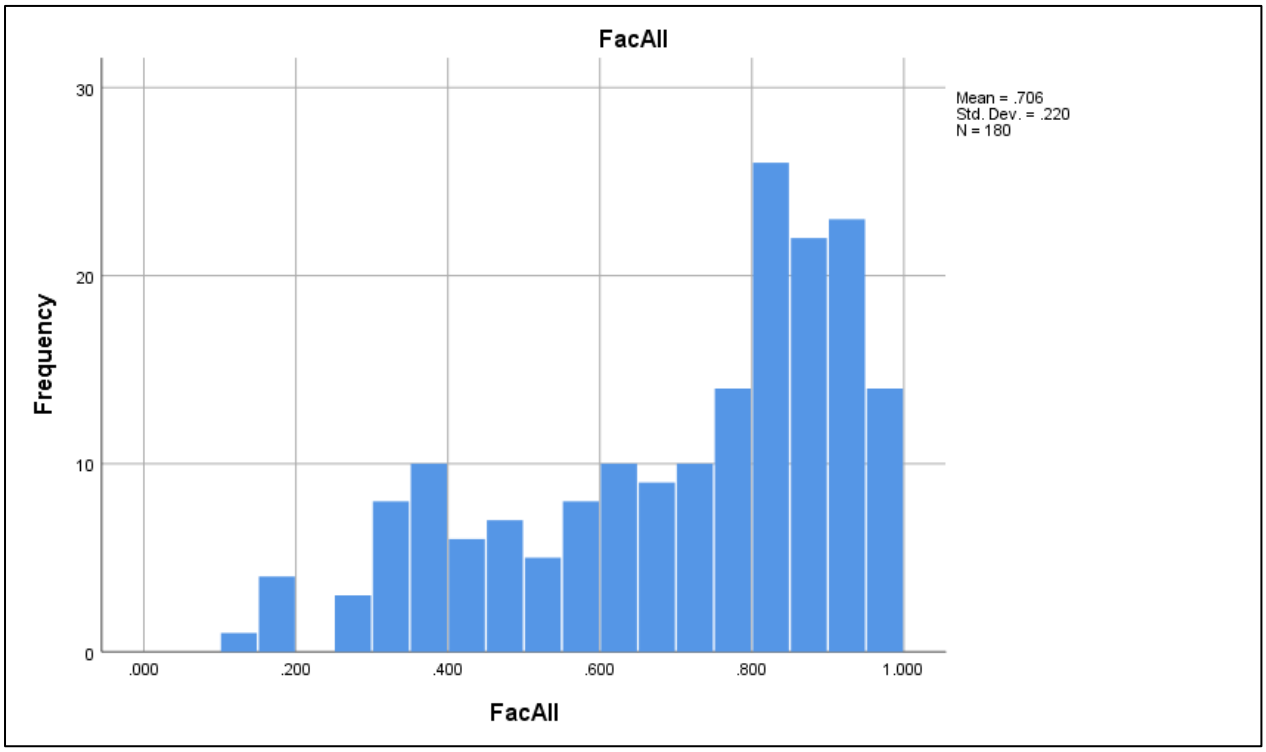

Figure 3: Distribution of item facilities (all candidates)

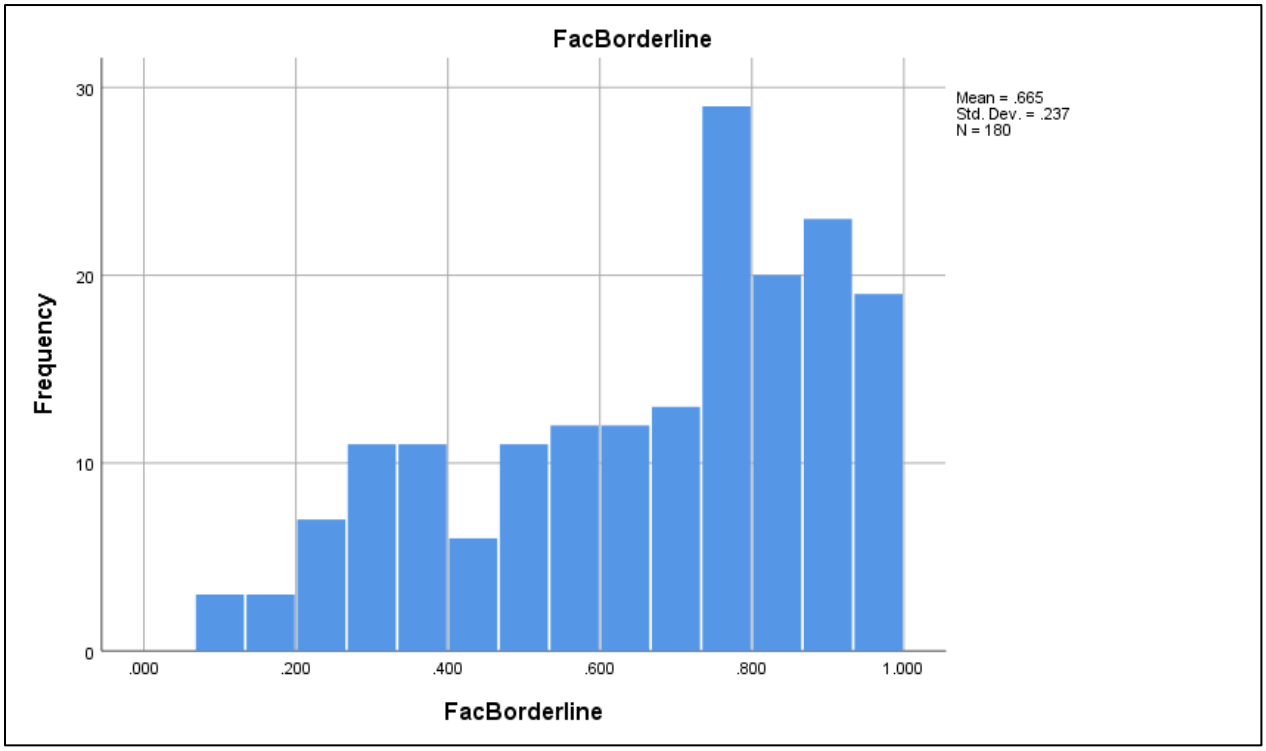

Figure 4: Distribution of item facilities for borderline candidates only

The differences in spread are also seen clearly in a comparative boxplot for these three scores (with median value also displayed):

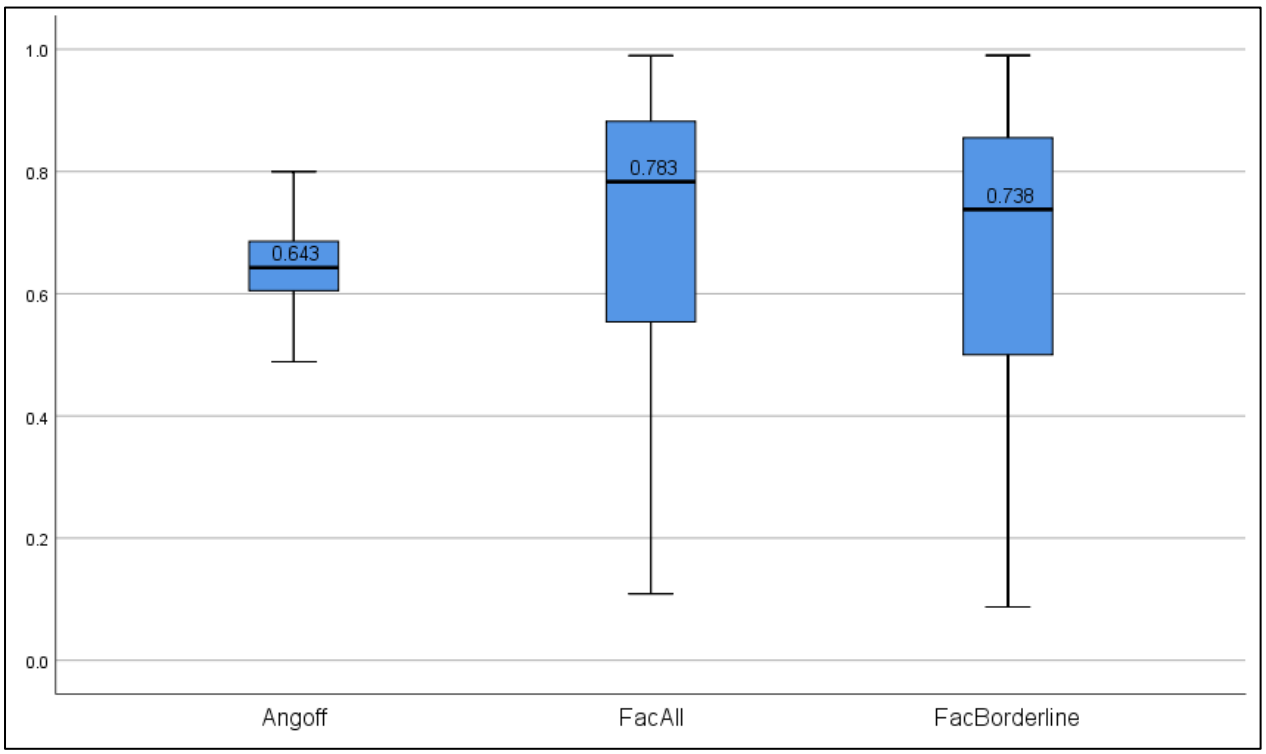

Figure 5: Comparative boxplot for the three item-level measures

In terms of the bi-variate relationships between the three measures, we can calculate correlations (all highly statistically significant):

	Facility All	Facility Borderline
Angoff	0.482	0.484
Facility All	1	0.992

Table 4: Pearson correlation between Angoff judgments and item facilities

Note, in Table 4, that the strength of the correlation between the two facilities is strong (r=0.992), and that the difference between their correlations with the Angoff judgments is small (r=0.482, r=0.484).

These insights are confirmed by the scatter matrix for the three item-level measures:

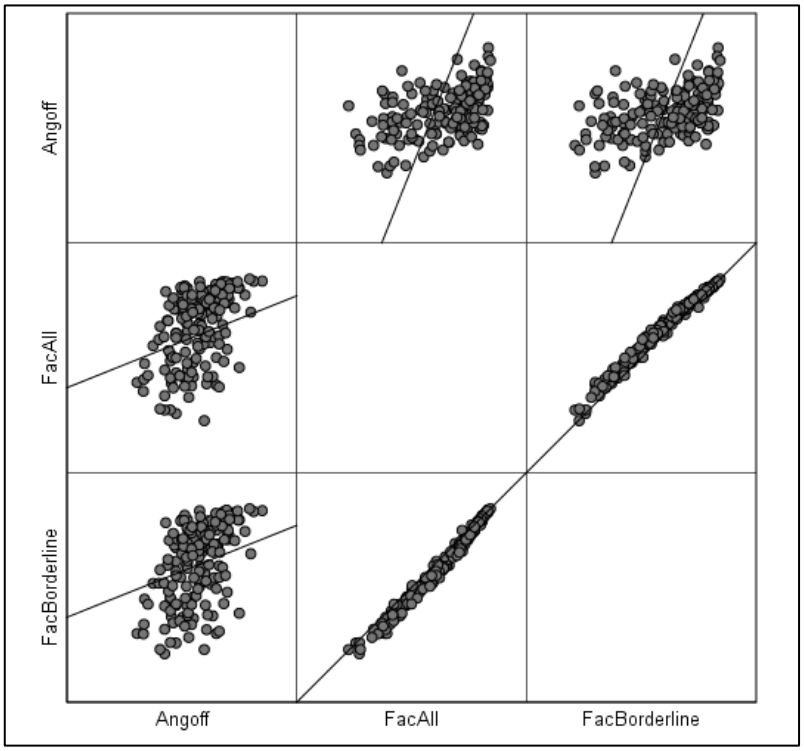

Figure 6: Matrix scatter plot for Angoff and item facility

The straight lines in Figure 6 are those of agreement between each pair of scores (i.e. y=x).

Perhaps the most relevant scatter graph shown in Figure 6 is that between Angoff and facility for the borderline group, which is shown in greater detail in Figure 7:

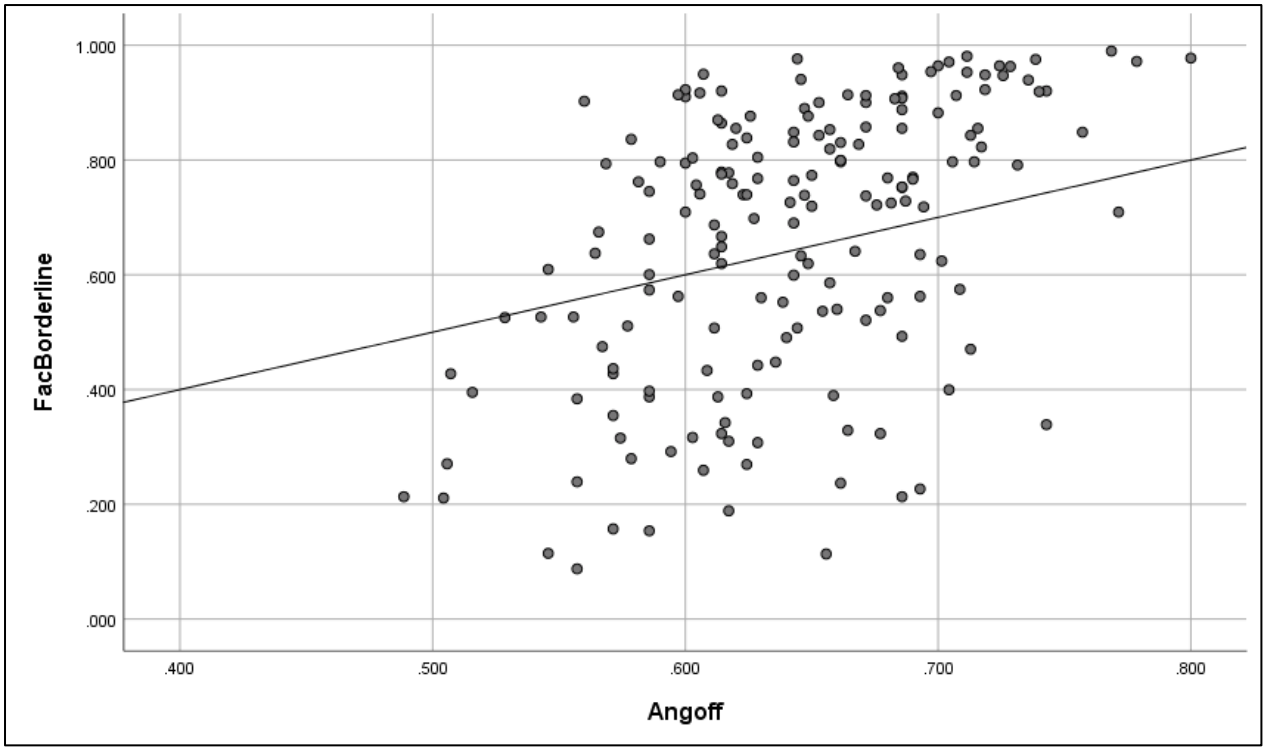

Figure 7: Scatter graph of Angoff versus item facility for borderline candidates

In Figure 7, those points above the line (n=110 out of 180, 61.1%) correspond to 'easy' items where the facility is higher than that expected in the Angoff judgements – in other words, those items that were easier than expected (and conversely for those items below the line – n=70, 39.9%). Again we see that actual item performance is far more spread out (vertically) than are the Angoff judgments (horizontally).

Another illuminating illustration of the relationship between these measures is the following (Figure 8), which, with an individual line for each item, pairs the Angoff judgment to the corresponding facility in the borderline group:

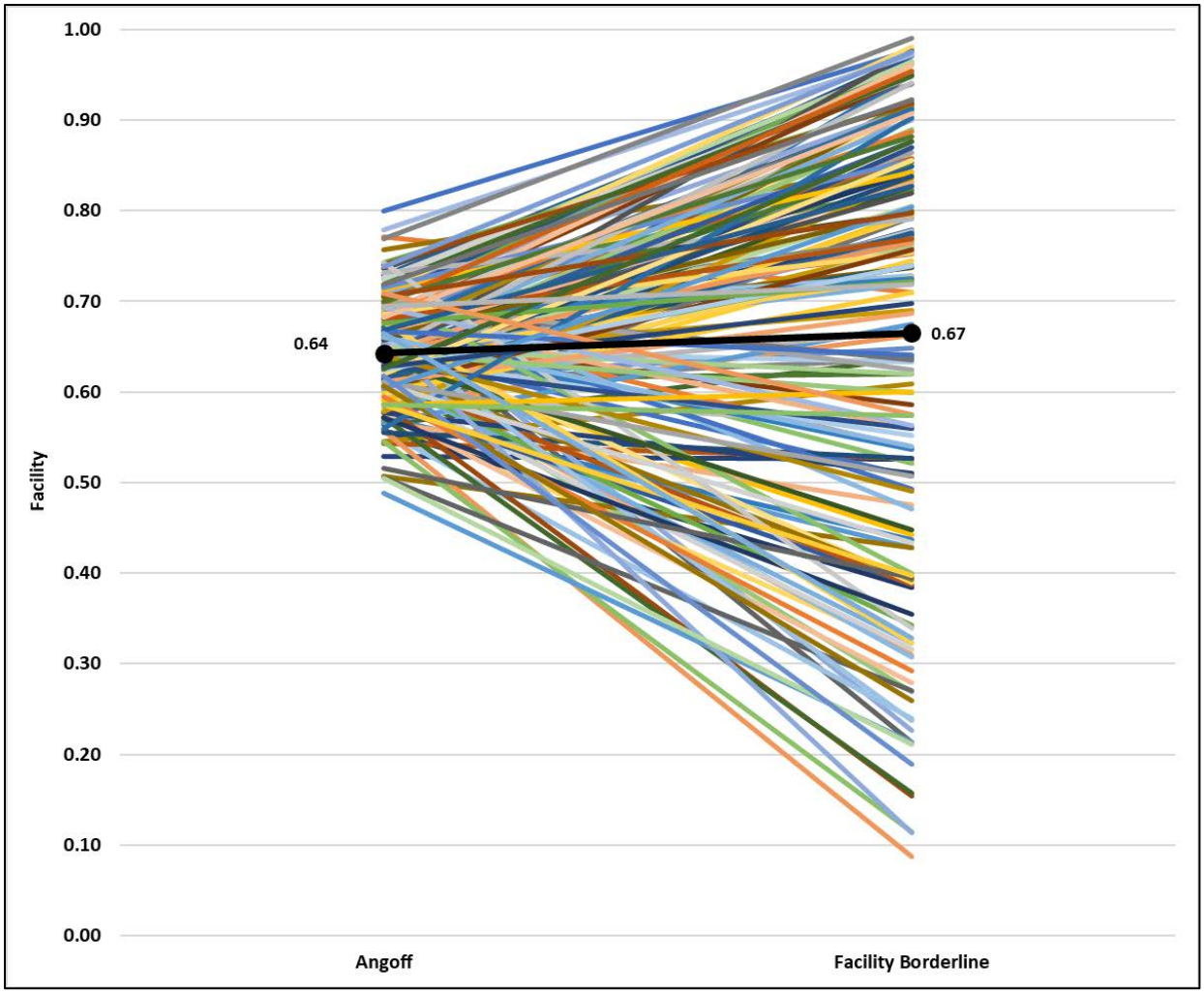

Figure 8: Line graph of Angoff versus facility in borderline group

The bold black line joining the two dots shows the mean of each set of measures.

Figure 8 shows clearly that the range of values for the actual facility in the borderline group is much wider than that reflected in the Angoff judgments. In general, harder items (towards the bottom of Figure 8) are harder than expected by the Angoff judges, and easier items (towards the top) are easier than expected.

However, it also shows that the two measures are close in mean value suggesting that differences between the Angoff judgments and the actual item performance in the borderline group typically balance out across the examination. This pattern of individual disconnect at the item level, but overall agreement at the test level, is common in a range of settings (Homer et al., 2012; Homer et al., 2019). It is also, however, largely an artefact of the circular nature of the definition of the borderline group in this analysis – which depends on the overall Angoff score, so there can be no independent verification in this analysis that the actual standard set is appropriate.

Summary of findings for RQ1

This analysis suggests that it might be worth reviewing the Angoff process and judgements. Angoff judges might be encouraged to use a wider range of expected facilities when making their judgments, although it is recognised that getting Angoff judgments 'correct' is very difficult in practice, particularly at the individual item level – see, for example, Clauser and

colleagues' work (2009). The recent addition of trainees to the Angoff panel is likely to bring an additional 'reality check' to the judgments made, and might well improve the alignment of these judgments with item performance. This will obviously be something to monitor as newly judged items work their way into PLAB1 examinations, and can then be compared with actual item performance.

RQ2 – Item performance over the course of PLAB1

Here we are interested in the extent to which items show a tendency to get harder, or are not attempted as the test progresses, particularly for the borderline group. This might indicate that weaker candidates are unduly time-pressured towards the end of the PLAB1 exams. We look at each of these issues in turn.

Item position versus item facility

Again, we focus mainly on March 2019 data. A scatter graph of item position (x) versus item facility (y), for all candidates and then separately for the borderline group, are shown below:

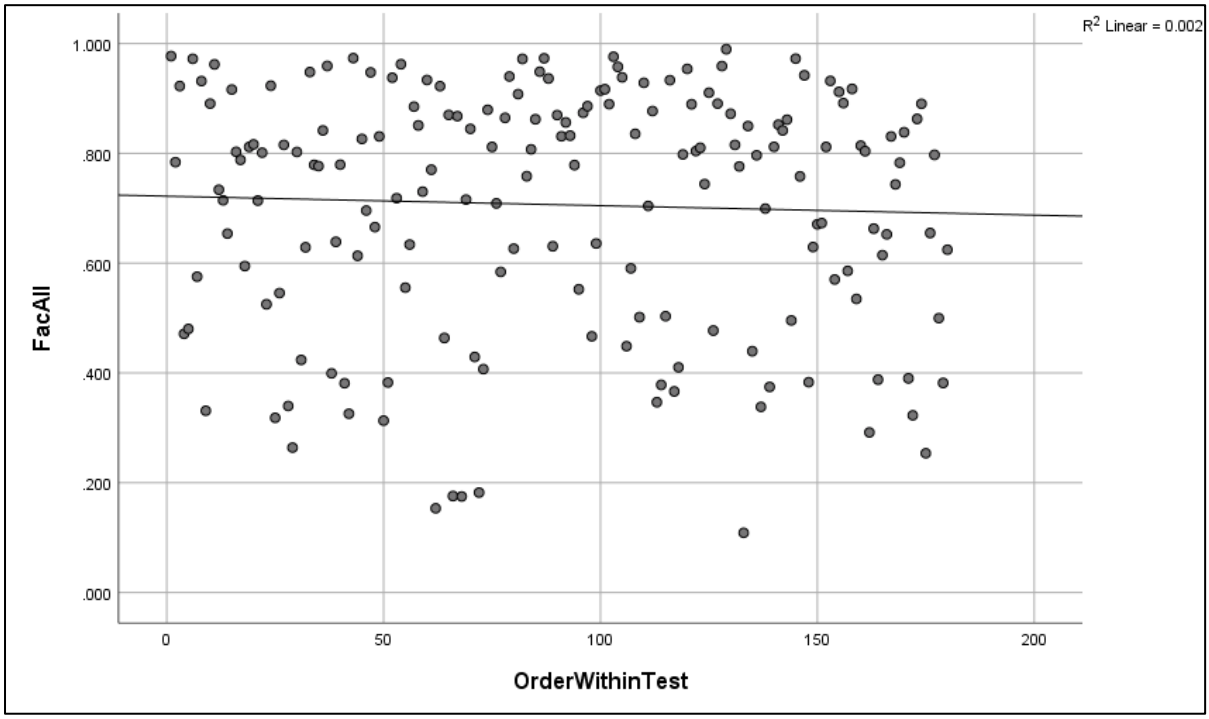

Figure 9: Item position versus facility (all candidates)

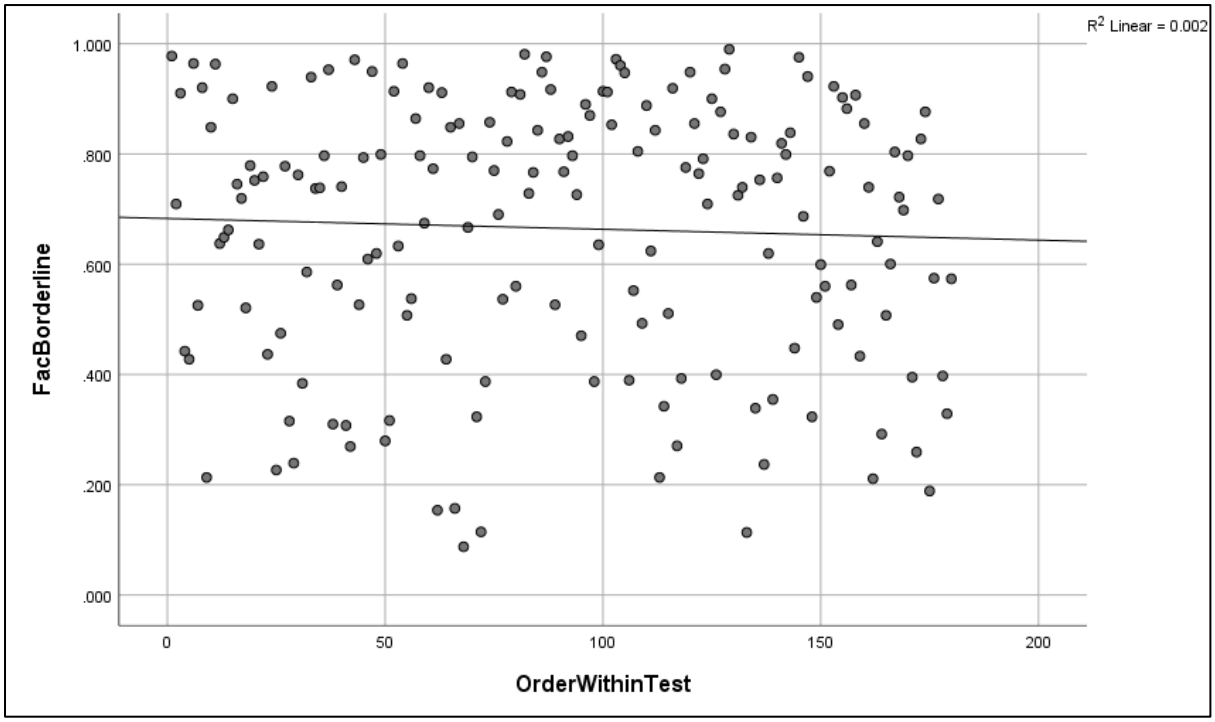

Figure 10: Item position versus facility (borderline candidates)

There is no apparent pattern in either case (the lines of best fit are very close to horizontal, with a very slight downward trend), and both correlations are not significantly different from zero (r=-0.041, r=-0.043 respectively). In other words, there is no real evidence in a decline in facility over the course of the March test. The results are entirely similar for the other three sittings. Given that items are positioned essentially at random in the tests, this analysis clearly suggests that item position is not a significant factor across the test. In other words, this analysis suggests that candidates, particularly those near the borderline, are not showing a significant decline in item performance over the duration of the test.

However a slightly different picture emerges if we split the items into deciles by item position (e.g. treat items 1-18 as a group, and 19-36 as another group and so on), and calculate the mean and median facilities for the borderline group within these bands. This allows more localised item position effects to be investigated.

We include all four sittings here as the results are a little different across these although broadly speaking there is no strong evidence of a serious decline in item-level performance across as each test proceeds – with, arguably, March being the exception to this:

Figure 11: Mean and median item facility by item position decile (March 2019, borderline candidates only)

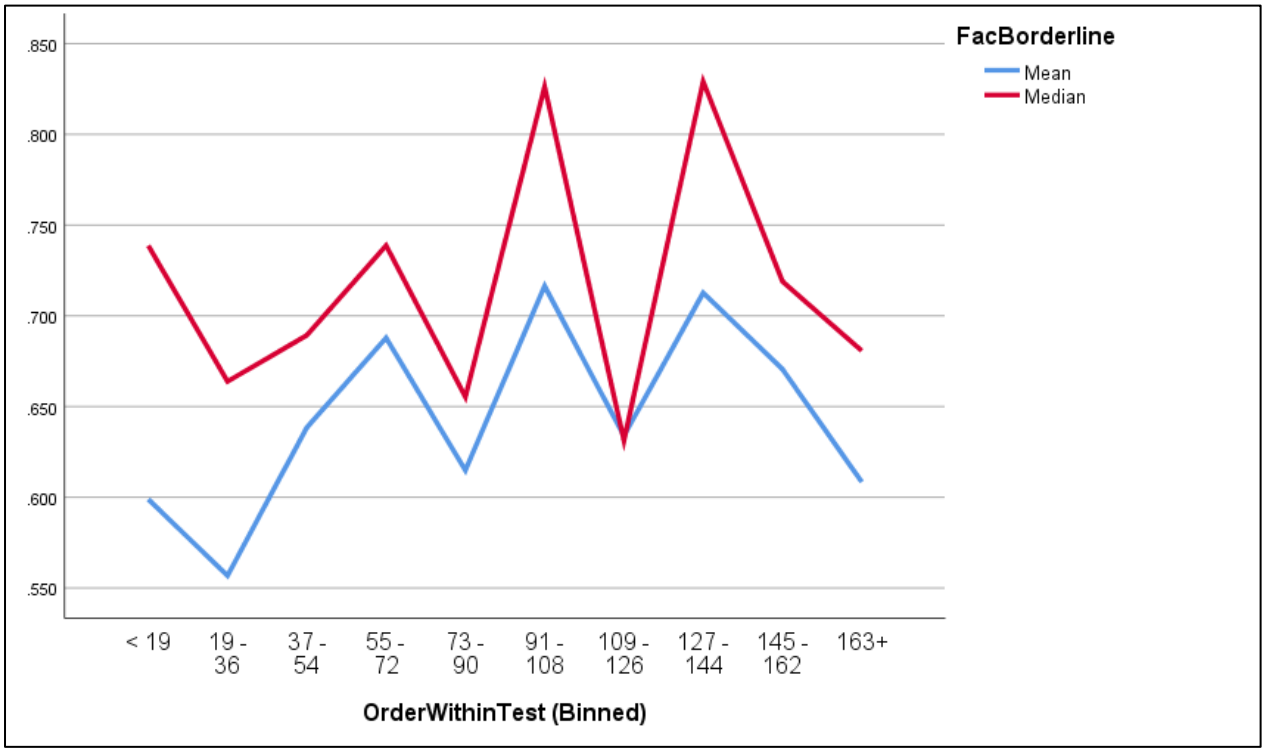

Figure 12: Mean and median item facility by item position decile (June 2019, borderline candidates only)

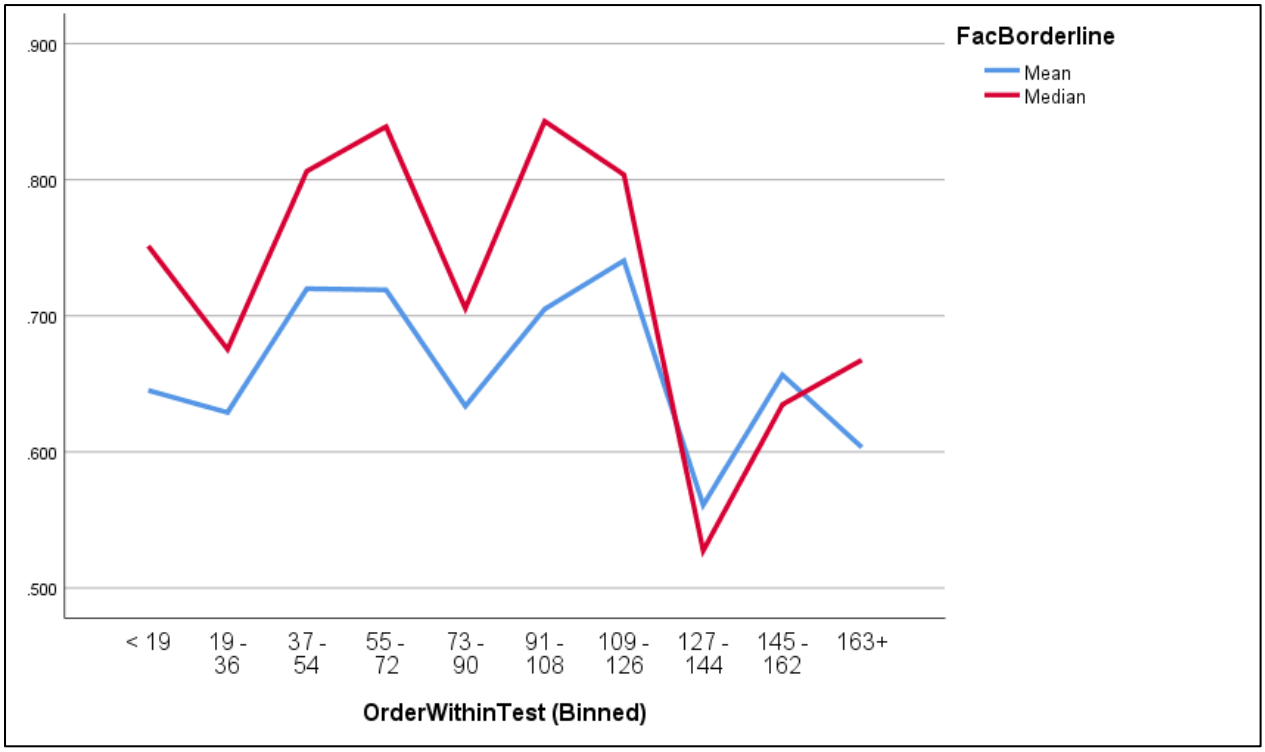

Figure 13: Mean and median item facility by item position decile (Sept. 2019, borderline candidates only)

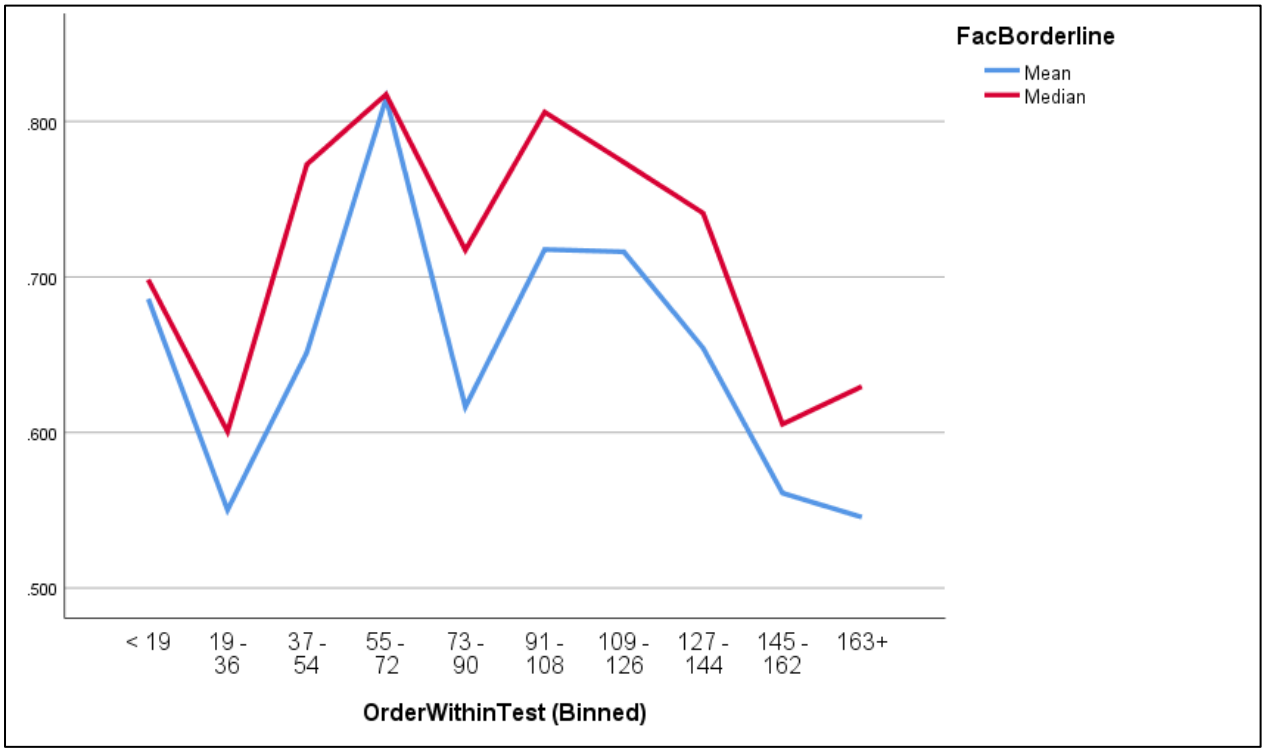

Figure 14: Mean and median item facility by item position decile (December 2019, borderline candidates only)

Item position versus item non-response

Another way of investigating whether borderline candidates are time-pressured towards the end of each examination is to count the proportion of them not responding to each item, and to see if this increases over the course of the exam. The scatter graphs that follow plot item position against proportion of non-response in the borderline group for that item. It is clear that typically there is very little non-response (generally less than 1% of this group), but that towards the end of the test this increase a little – but still remains a relatively low percentage of candidates (≈2-3% at most).

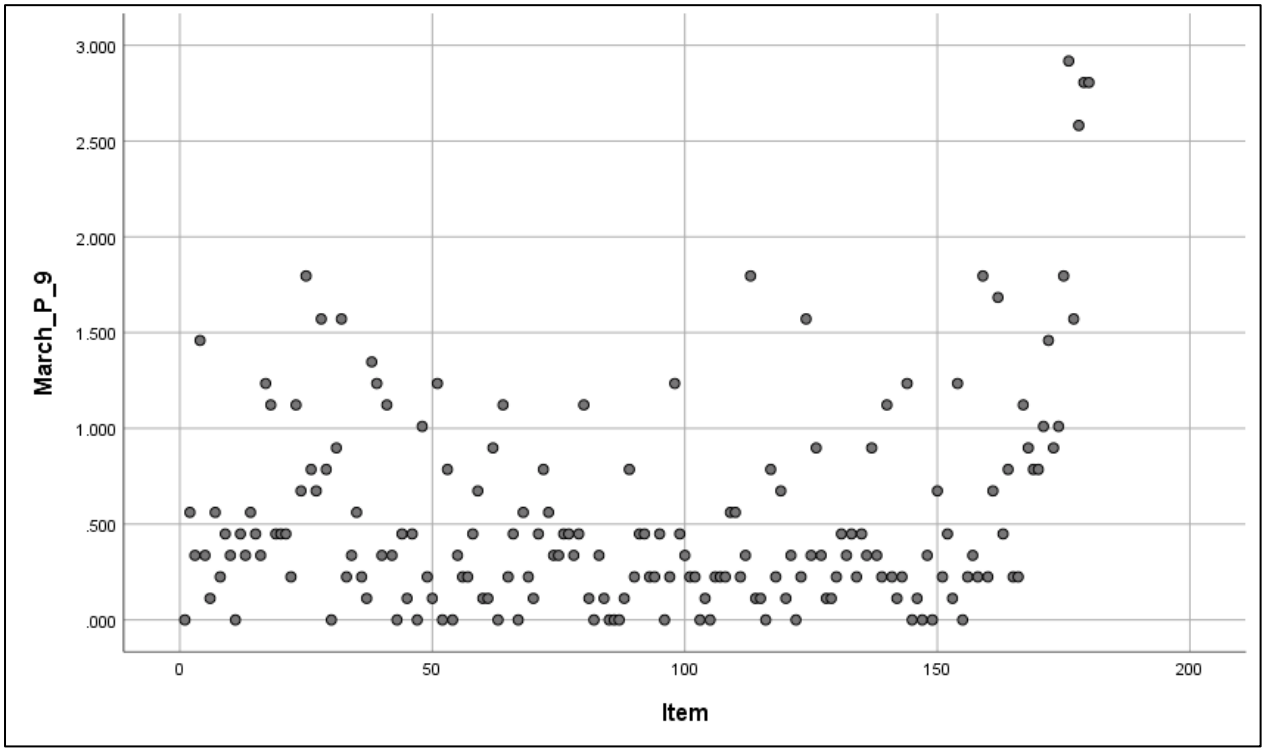

Figure 15: Item position versus proportion of item non-response in the borderline group (March)

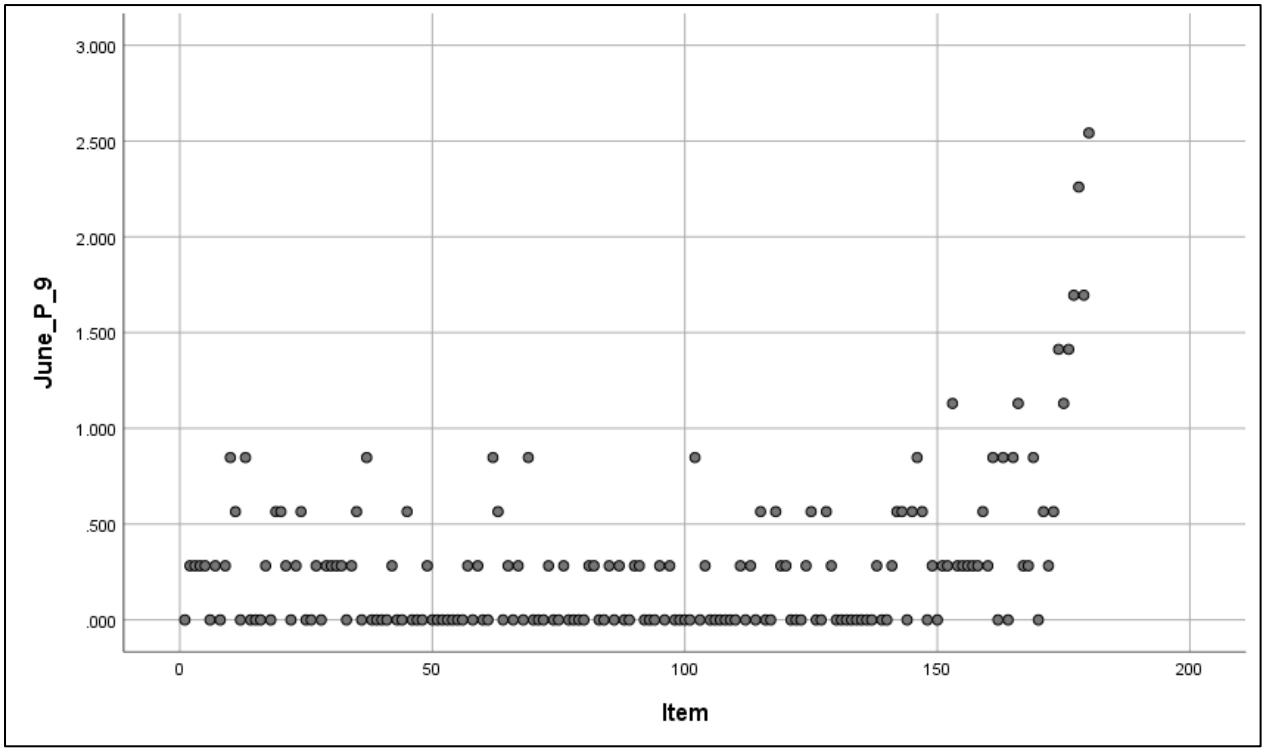

Figure 16: Item position versus proportion of item non-response in the borderline group (June)

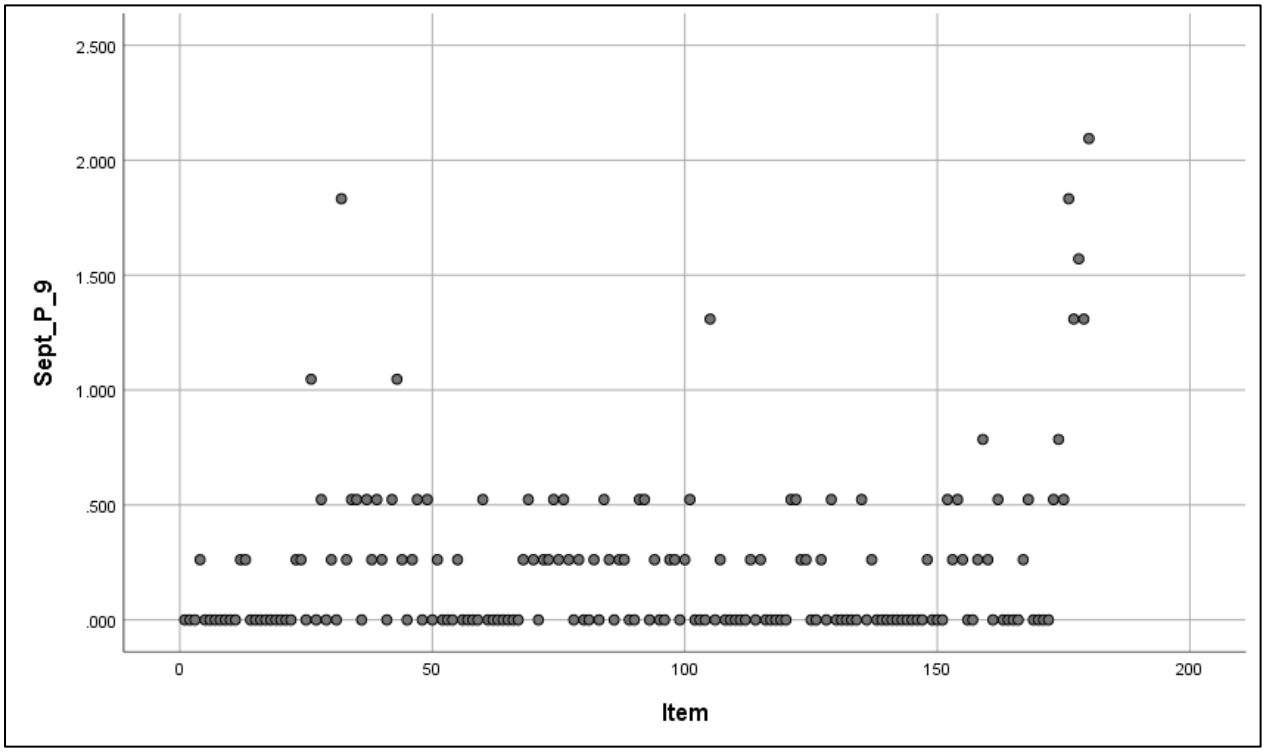

Figure 17: Item position versus proportion of item non-response in the borderline group (September)

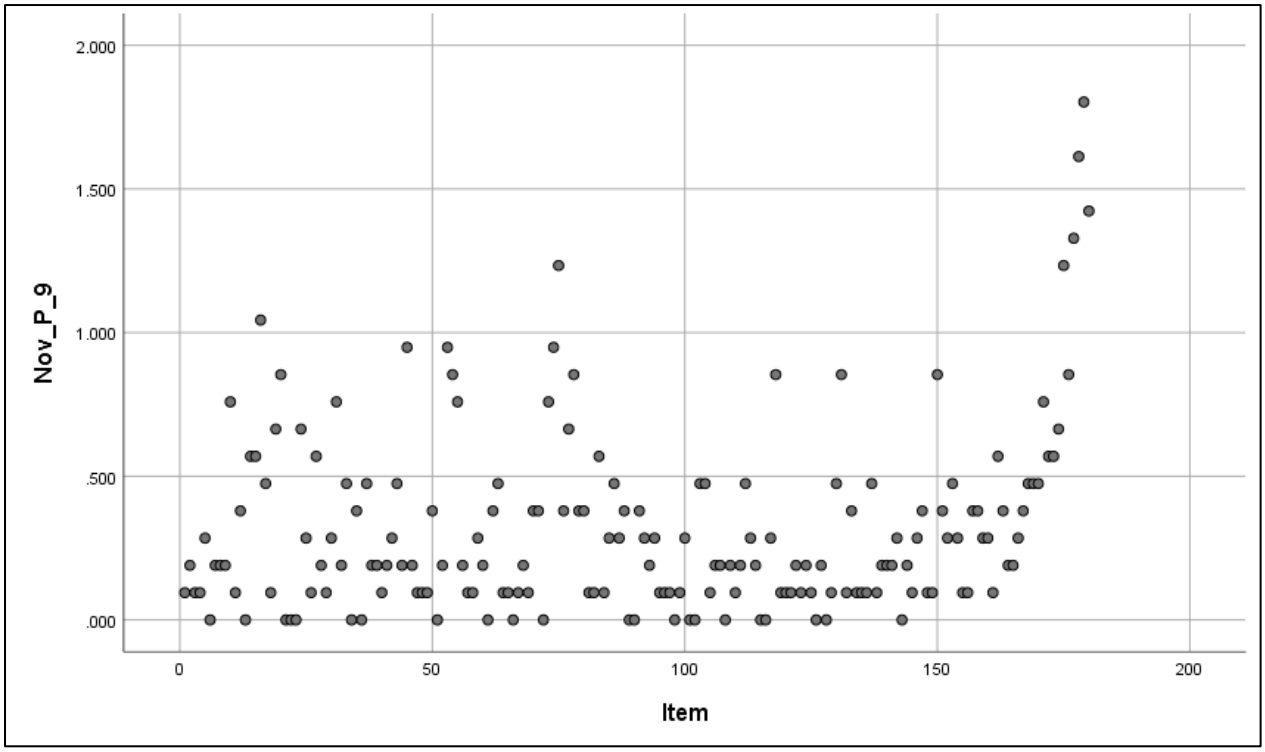

Figure 18: Item position versus proportion of item non-response in the borderline group (November)

Summary of findings for RQ2

Generally speaking, there is no evidence in this data to suggest that weaker candidates are unduly suffering significant time-pressure towards the end of the test. The current test length of 180 items, with 1 minute per item, therefore seems perfectly adequate and fair, even for borderline and/or weaker candidates.

RQ3 – Relationship between PLAB1 and PLAB2 scores

We want to get a measure of the predictive validity of PLAB1 for PLAB2 performance – in other words, the correlation between these two scores. However, there are several methodological complications when calculating this including multiple attempts at PLAB1 (and PLAB2), missing data in PLAB2 for those that do not pass PLAB1, and measurement error in both measures.

Perhaps the purest measure of the correlation between performances on the two exams is where we take only first attempts at PLAB1 and measure the correlation with PLAB2 first attempt and correct for 'missing' values on PLAB2 via imputation for those that failed on this first attempt[3]. For each candidate, we do not use raw scores as the measure of performance since the standards in the test (and the test length on occasion) vary. Instead we calculate

[3] We use the expectation-maximization algorithm as implemented in SPSS to do this – assuming a normal distribution of PLAB2 scores. We also require PLAB2 to be missing – there are 76 cases where later attempts at PLAB1 in 2019 resulted in an actual PLAB2 score based on a later attempt at PLAB1 – these have been removed and treated as missing in this analysis.

the percentage score above the pass mark for each of PLAB1 and PLAB2 – this then adjusts to an extent for any variation in test difficulty and test length[4].

The first two graphs show histograms of these two performance measures separately.

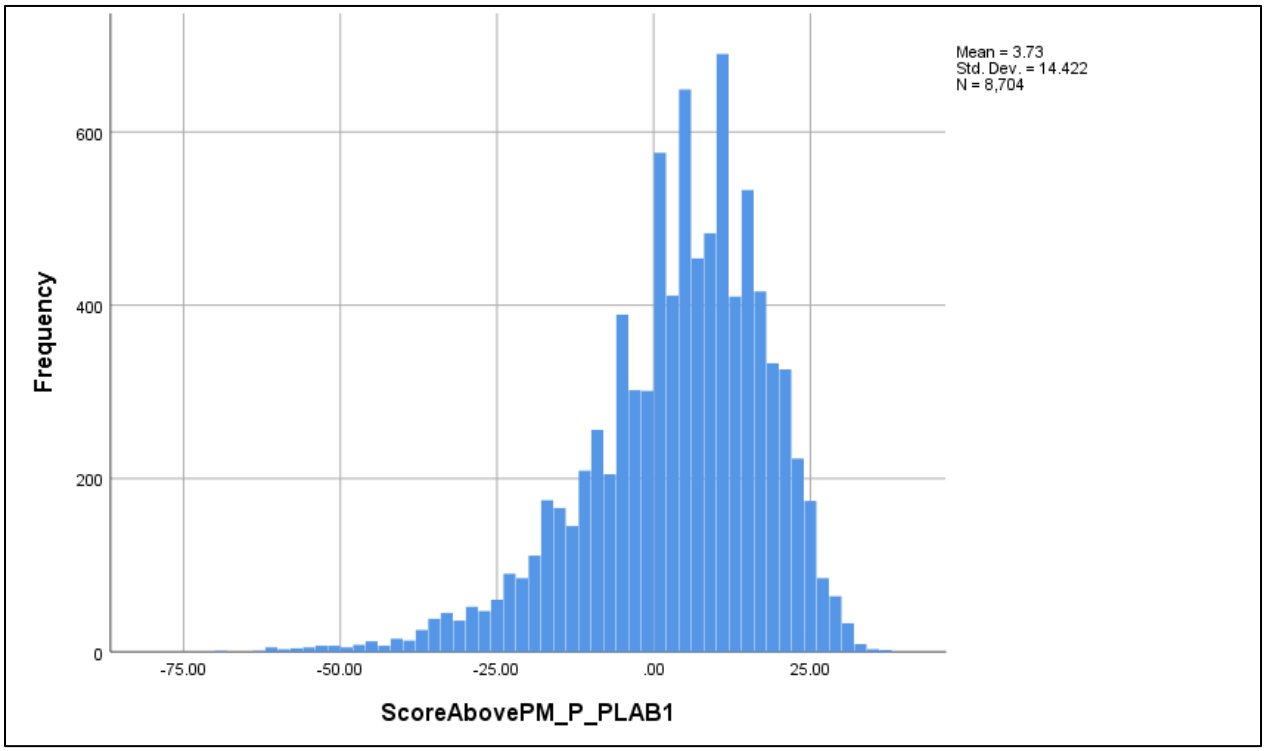

Figure 19: Distribution of PLAB1 first attempt scores (% above pass mark)

[4] For convenience, we use the standard set passing score as the baseline for this percentage calculation. It might be preferred to use the length of the scale (i.e. total marks available in each test) but due to limited test level data on supressed items/stations, this work around is convenient and will make little difference overall to the key results.

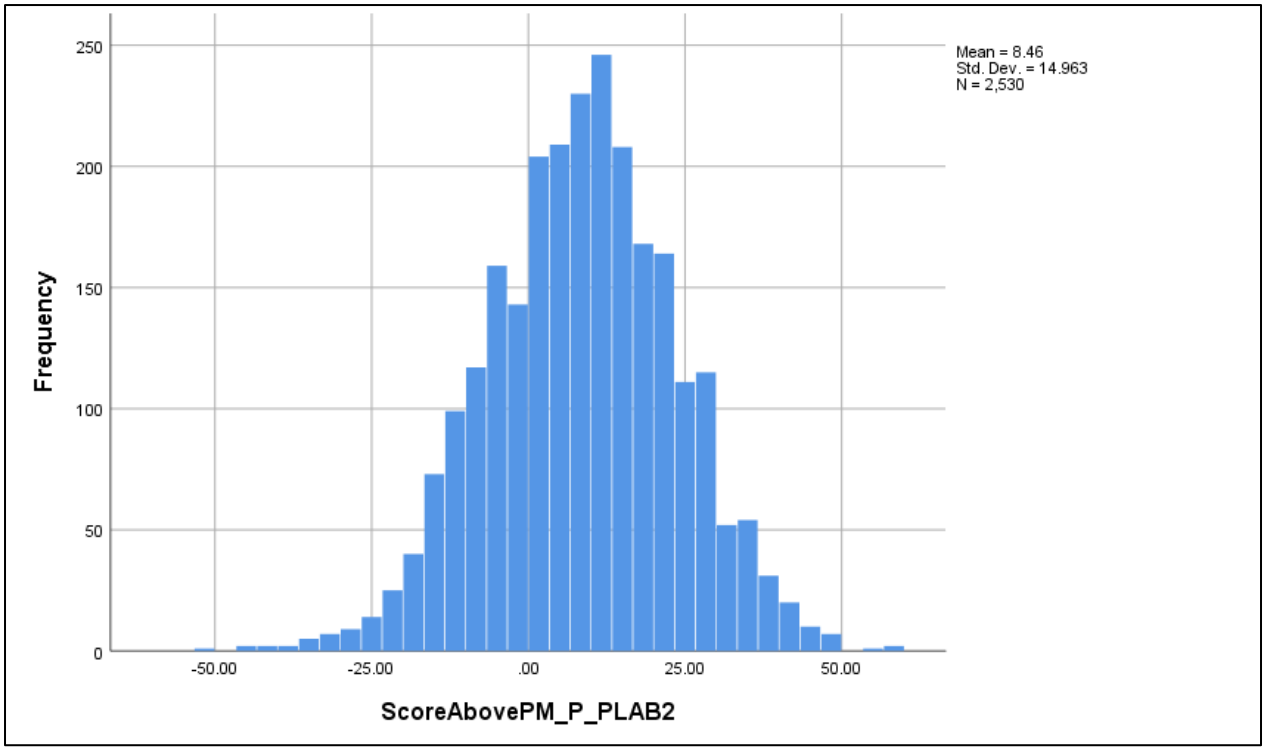

Figure 20: Distribution of PLAB2 first attempt scores (% above pass mark)

We can see there is a lot more PLAB1 data (n=8,704), than PLAB2 (n=2,530) – and we will deal with the 'missing' PLAB2 data shortly.

When we plot the joint distribution via a scatter graph, we see in Figure 21 the key statistical problem – the 'missing' PLAB2 scores for those that failed PLAB1 and have yet to sit PLAB2. This restricts the data range for PLAB1, which in turn will weaken the correlation between the two scores. Hence the need to impute the 'missing' PLAB2 scores to get a better estimate of the 'true' correlation between them.

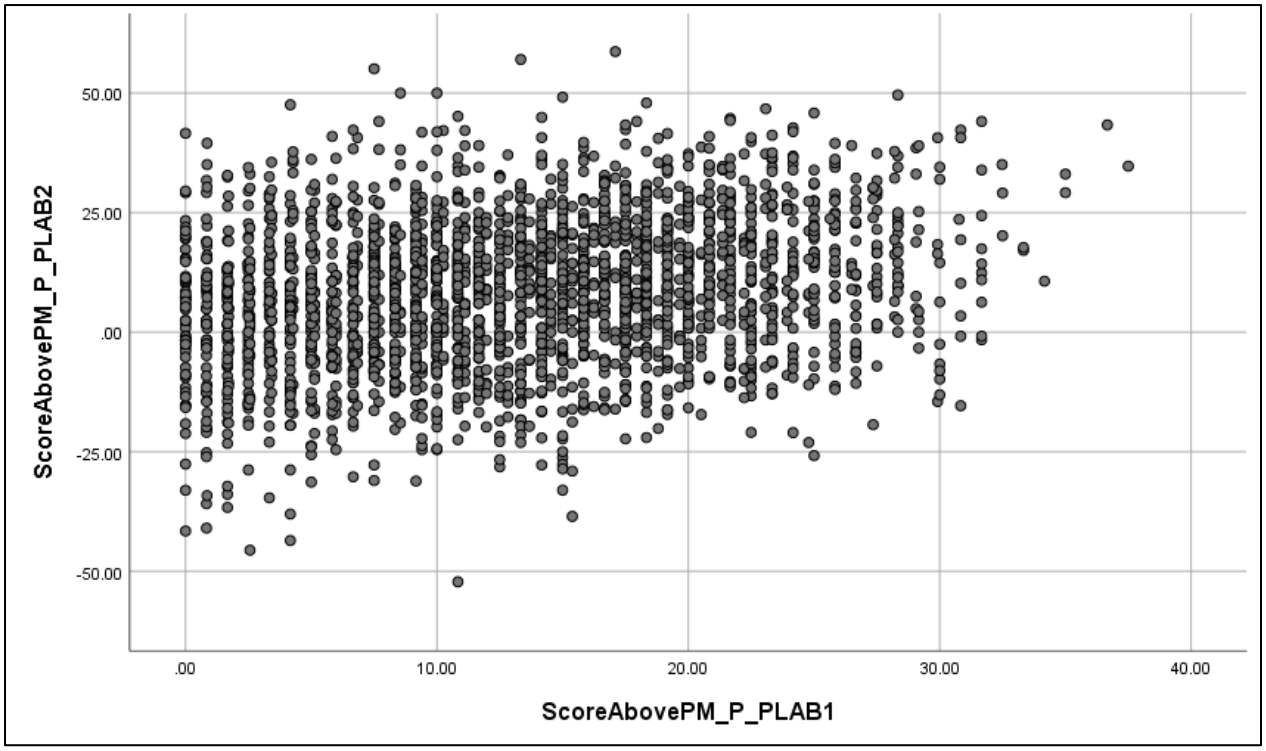

Figure 21: Percentage scores above pass mark – PLAB1 versus PLAB2

Using the data shown in Figure 21, the raw correlation between the two percentages is r=0.259 (p<0.001, n=2,530) – but this is very likely an underestimate for two reasons – one, the missing data problem described above, and two, the lack of perfect reliability in the scores – measurement error in any pair of scores will weaken the correlation between them (Trafimow, 2016). We correct for these in turn.

The correlation using imputed data for PLAB2 increases to 0.432 (n=8,704), but this analysis should probably be regarded as indicative rather than completely authoritative as the required assumption regarding data missing at random is violated in the data (e.g. the missing PLAB2 scores are for weaker candidates, rather than from a random sample of them).

If we then also correct for measurement error (i.e. imperfect reliability, alpha=0.89 for PLAB1 and 0.74 for PLAB2 using typical values) this correlation increases to r=0.532.

This is a 'large' correlation according to the usual guidelines (Cohen, 1988), and is in line with, but a bit larger, than other correlations between knowledge and performance tests evident in a quick survey of the relevant literature (Dong et al., 2012; Park et al., 2015). However, these latter papers do not have the missing data issue, and sometimes do not correct for the imperfect reliability of the tests so direct comparisons with their findings are difficult.

Summary of findings for RQ3

This analysis gives confidence that performance in PLAB1 does predict to an extent PLAB2, but, as we might hope, strong applied knowledge performance as evidenced in PLAB1 is no guarantee of automatic success in PLAB2.

RQ4 – Impact of the shift to two circuits in PLAB2

As the demand for PLAB2 has grown, the exam has moved from a single circuit to two parallel circuits (in August 2019). This growth in the size of the candidate pool, and the subsequent need for two examiners per station instead of one, raises a number of key questions around the assessment quality of the exam under its new format, and comparisons with the simpler, single circuit model of the past.

We begin this analysis with an investigation into overall exam reliability before and after this change, and then go on to compare station level metrics, and finally metrics across the parallel circuits in the new format.

Overall reliability – one circuit versus two

In 2019, there were 174 test administrations, of which 64 (37%) were with two circuits (Table 5).

Circuits	Mean	Median	Minimum	Maximum	SD	N
Single circuit	0.75	0.77	0.42	0.87	0.08	110
Two circuits	0.72	0.73	0.52	0.83	0.06	64
Total	0.74	0.75	0.42	0.87	0.07	174

Table 5: Comparison of descriptives for reliability (alpha) (1 vs 2 circuits)

There is a slight decline in mean reliability when moving to two circuits (from 0.75 to 0.72; p=0.007, Cohen's d=0.45) - a statistically significant change in mean reliability with a 'moderate' effect size according to the usual guidelines (Cohen, 1988).

This decline in reliability is probably expected given that any systematic differences between examiner scoring in the same station in the same administration will generally add to measurement error, and hence lower reliability.

The distributions of alpha values across the 174 administrations are shown in the following histograms – blue for single circuit (left), and red for two circuits (right).

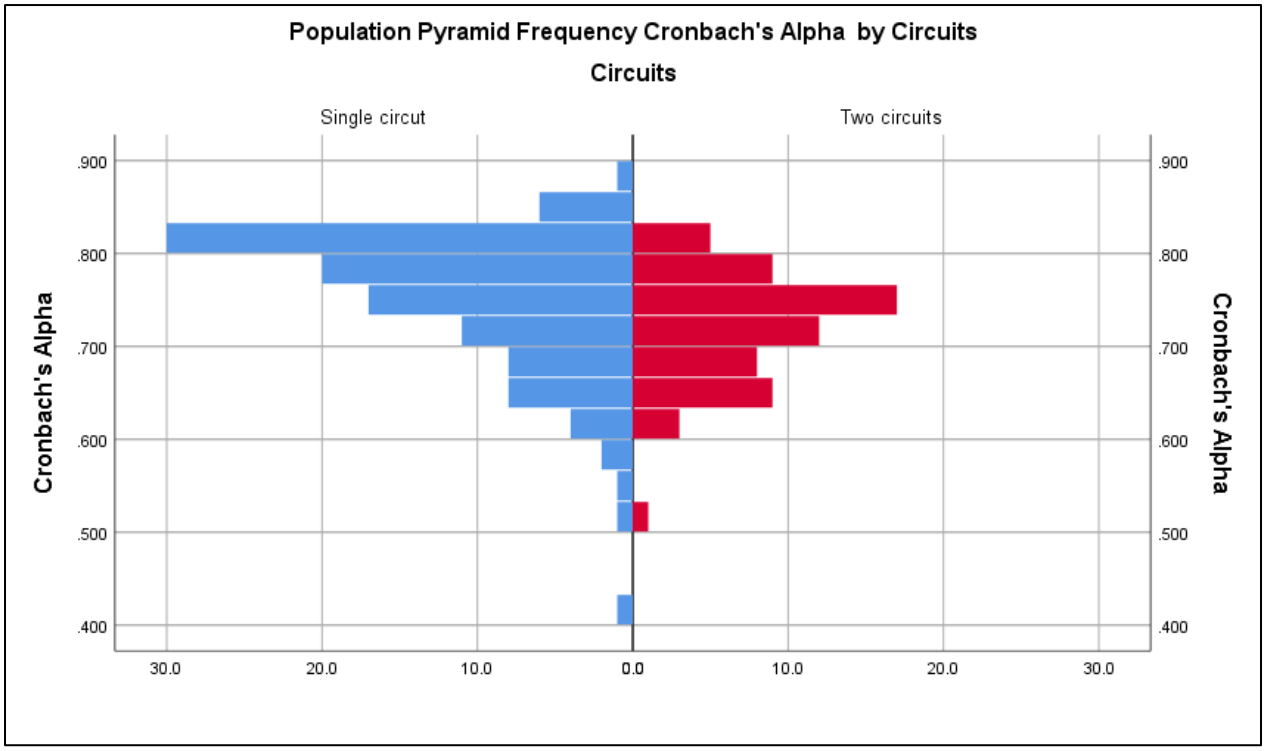

Figure 22: Comparative histogram of reliability (single vs two circuits)

Station level metrics – one circuit versus two

At the station level, a comparison of various metrics generally indicates small differences on average. For example, mean R-squared is slightly lower (changing from 0.76 to 0.74; $p<0.001$; Cohen's $d=0.21$) – again, this is probably what we would expect – any underlying difference between patterns of scoring by examiners in the same station (but parallel circuit) will lower R-squared.

Circuits	Statistic	Facility	Slope	Intercept	Cut score	R-squared
Single circuit	Mean	0.67	2.10	3.57	6.08	0.76
	Median	0.69	2.08	3.58	6.04	0.79
	Minimum	0.03	-2.54	-2.29	2.43	0.04
	Maximum	1.00	5.19	10.11	9.11	0.98
	SD	0.19	0.53	1.07	0.79	0.12
	N	1,953	1,953	1,953	1,953	1,953
Two circuits	Mean	0.65	2.17	3.61	6.22	0.74
	Median	0.67	2.17	3.61	6.18	0.76
	Minimum	0.01	-0.33	0.53	4.58	0.03
	Maximum	1.00	3.71	7.22	8.84	0.94
	SD	0.17	0.40	0.85	0.61	0.11
	N	1,136	1,136	1,136	1,136	1,136
Total	Mean	0.66	2.12	3.59	6.13	0.75
	Median	0.69	2.12	3.59	6.11	0.78
	Minimum	0.01	-2.54	-2.29	2.43	0.03
	Maximum	1.00	5.19	10.11	9.11	0.98
	SD	0.18	0.49	1.00	0.73	0.12
	N	3,089	3,089	3,089	3,089	3,089

Table 6: Comparison of station level metrics (single vs two circuits)

Comparison across blue and green circuits

The advent of two circuits allows for a detailed comparison across them of various station level metrics. There is very little published work on this (at least, to my knowledge) so the analysis here provides a bench mark for later investigations.

We begin with ANOVA R-squared values. This metric gives a measure of the differences in station-level domain scores across the two circuits – assuming candidates are randomised to circuits - and is useful as a proxy for differences in scoring across examiners in parallel circuits. Small values indicate similar scoring across circuits for the same station, and the standard acceptable threshold is usually 0.4 or 0.5 (Pell et al., 2010) .

N	Minimum	Maximum	Percentiles				
			5	25	50	75	95
1,136	0.00	0.87	0.00	0.02	0.07	0.19	0.44

Table 7: Descriptive summary of ANOVA R-squared values across circuits

We see that most values lie within the acceptable range, although these guidelines were original developed in undergraduate settings with a much larger number of candidates (e.g. n≈250 or more), and hence parallel circuits (e.g. 20 or more).

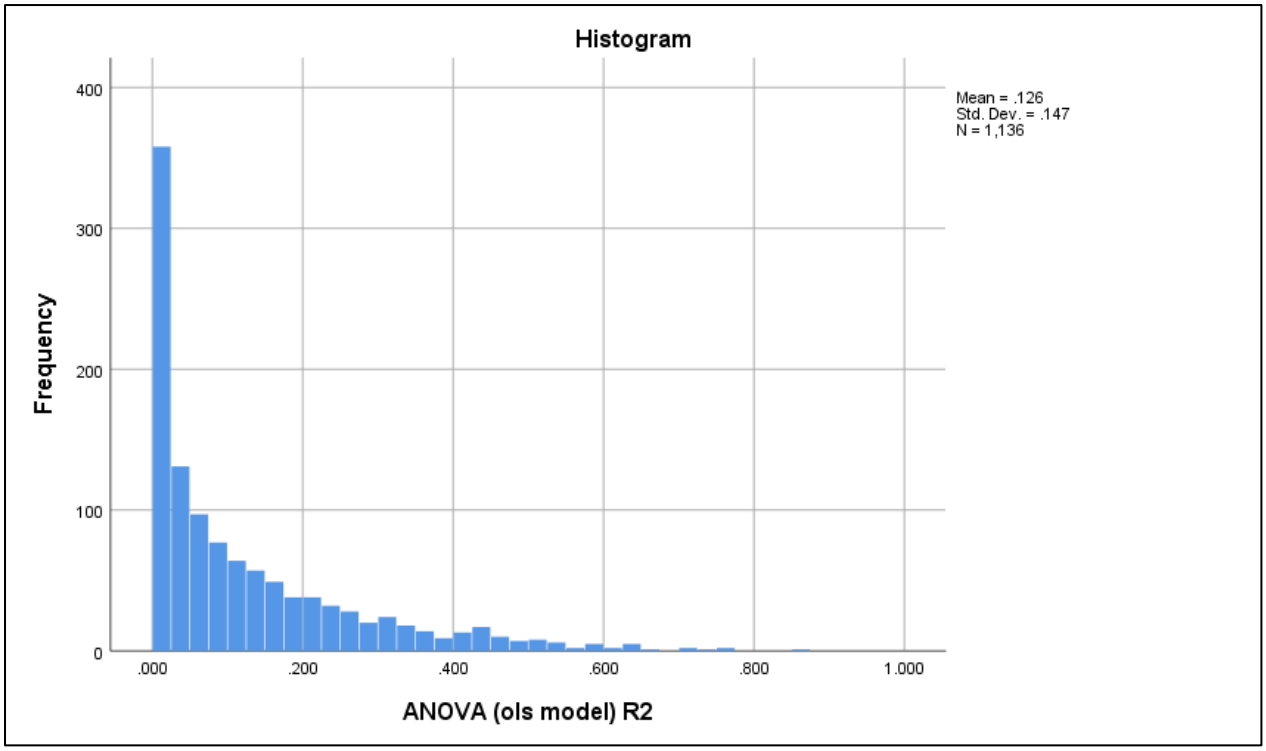

Figure 23: Histogram of ANOVA R-squared values across circuits

The distribution in Figure 23 is highly positively skewed with the vast majority of values well with the acceptable range.

The outlier station with the largest ANOVA R-squared (value 0.87) has a borderline regression graph as follows (blue line for blue circuit, green for green, and black for combined):

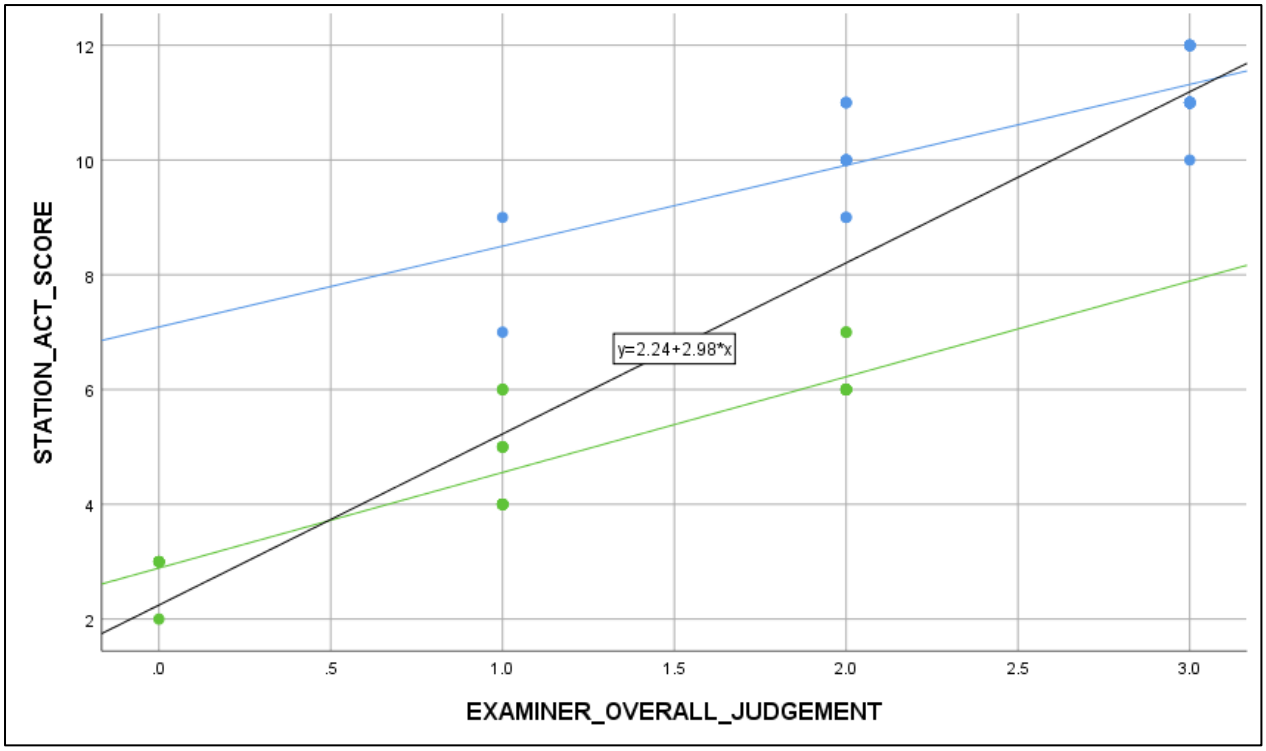

Figure 24: Station[5] with largest ANOVA R-squared metric

One can see that the two examiners are quite different in their scoring – the one on the green circuit scores systematically lower for the same global grade, and also is not giving as higher grades in general. We must remember that the two sets of candidates are, of course, different but the data here is suggestive of relatively hawkish behaviour for green (or the opposite for blue). This same station has been administered six times since two circuits were introduced and this high value of the ANOVA R-squared is very much an outlier (median value=0.03) – so this is suggestive of aberrant examiner behaviour in this instance, rather than a general problem with the station. We should also remember that this is the worst case in over 1,000 station administrations since the move to two circuits.

We now move onto other metrics, investigating how much the slope varies for the same station in the same exam across circuits; similarly intercept, cut-score and R-squared (under borderline regression).

[5] 4988: Women attending an appointment with a personal issue

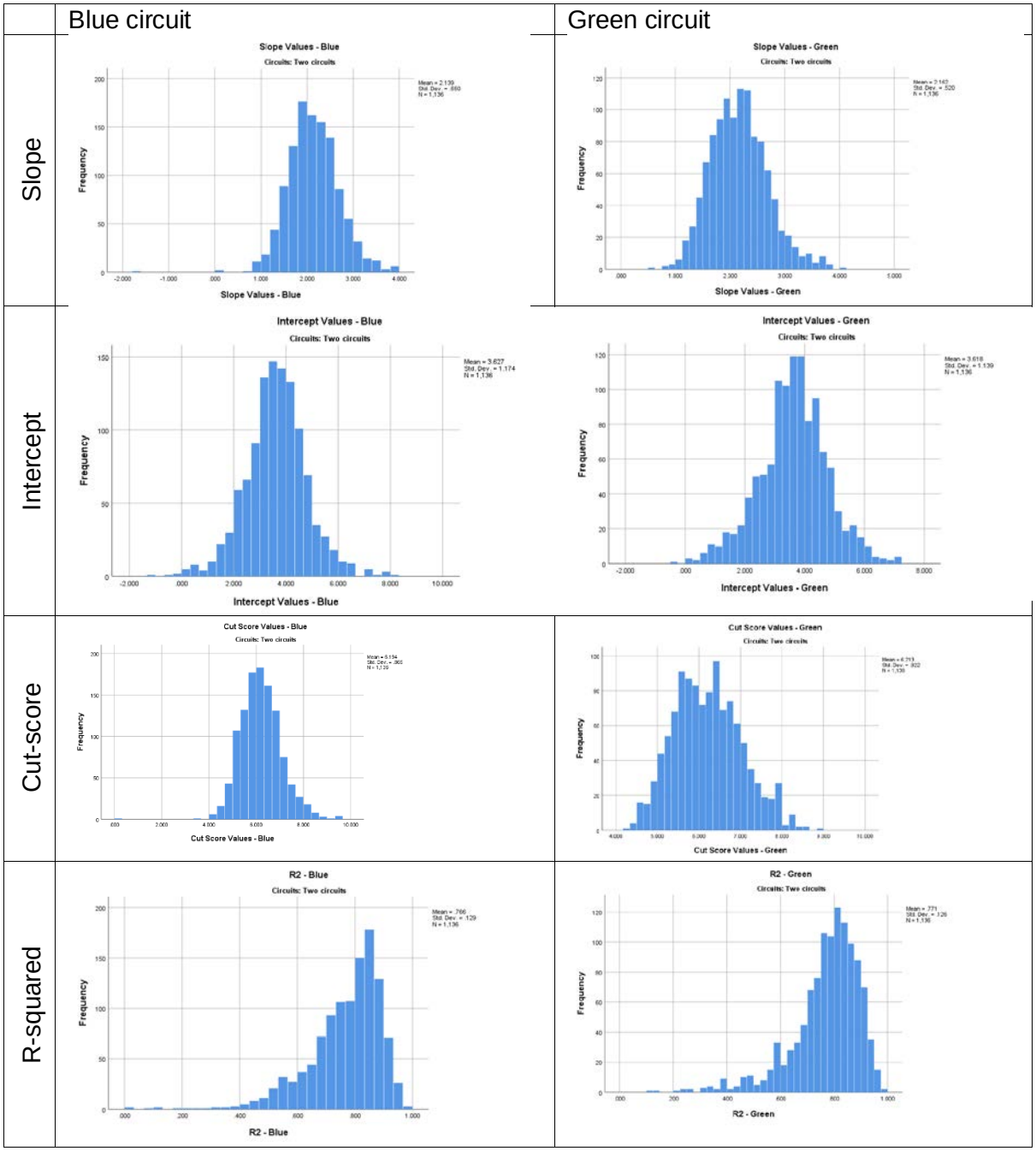

Figure 25: Histograms comparing metrics across parallel circuits

If we test the difference between blue and green circuits on these metrics (paired t-test), we find non-significant results overall – with mean difference as a percentage of the scale very small (<0.5% in all cases):

Metric	Mean difference	Scale	Mean difference %	SD	Std. Error Mean	t	df	p-value
Slope	-0.023	0-12	-0.19	0.735	0.022	-1.06	1135	0.29
Intercept	0.009	0-12	0.07	1.558	0.046	0.19	1135	0.85
Cut score	-0.019	0-12	-0.16	1.127	0.033	-0.56	1135	0.58
R-squared	-0.004	0-1	-0.44	0.171	0.005	-0.87	1135	0.38

Table 8: Paired t-tests comparing metrics in blue and green circuits

Note that the actual cut-score combines the data across the two circuits in the borderline regression calculation. Calculating this within the circuit is for illustrative purposes only – in order to investigate different patterns of scoring.

Table 8 indicates, as we might hope, that there is no evidence that circuits are behaving differently across the set of 2019 administrations when taken as a whole.

However, if we focus on differences in cut-scores (as perhaps the most important metric; see the third row in Table 8), we obtain the following summary statistics and histogram when this is expressed as a percentage of the total domain score scale in the station (0-12):

N	Minimum	Maximum	Percentiles				
			5	25	50	75	95
1,136	-42.4	33.8	-15.7	-6.3	-0.1	5.8	15.6

Table 9: Descriptive summary of percentage difference in cut-scores across circuits

There are some stations where the cut-scores would be very different between the two circuits, but the vast majority are within 15% of each other – (5^{th}, 95^{th} percentile) = (-15.7,15.6). As stated above, there is little evidence in the literature of how these metrics vary across circuits, so the extent to which these statistics are satisfactory is unknown, but initial thoughts are that they seem quite reasonable in the vast majority of cases.

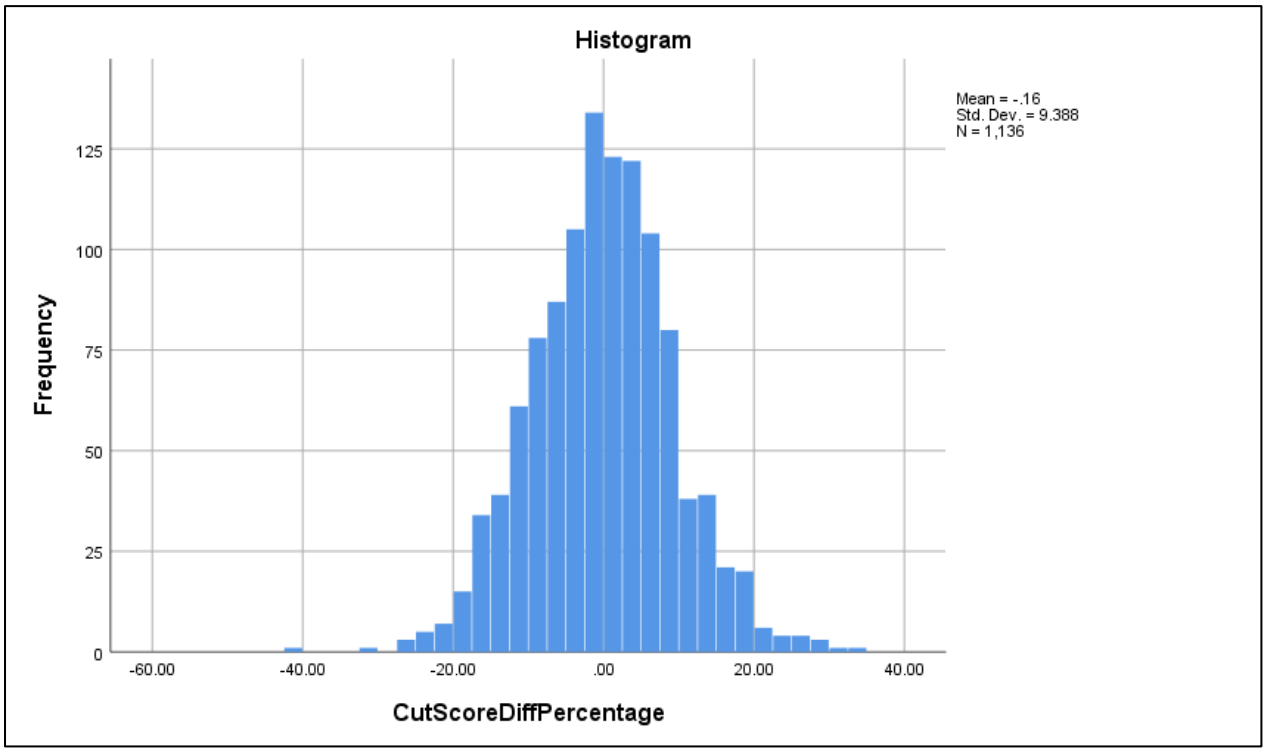

Figure 26: Histogram of percentage difference in cut-scores across circuits

Summary of findings for RQ4

Overall the analysis for this research question indicates that the move to two circuits has not led to any significant problems in terms of assessment quality and candidate outcomes. Much of the analysis presented here concerning differences in borderline regression metrics

and (hypothetical) cut-scores across parallel circuits is not covered in the literature in great detail, so this work is an initial attempt at filling that gap.

RQ5 – Candidate performance morning versus afternoon in PLAB2

For this RQ, we want to investigate if there is evidence of systematically different examiner behaviour between the morning and afternoon sessions of PLAB2, which might indicate problems of examiner fatigue as the day progresses.

In terms of variation in scores across morning and afternoon, we find that there is very little difference in means either at the domain, total station score, total exam score, global grade or (hypothetical) cut-score required[6]:

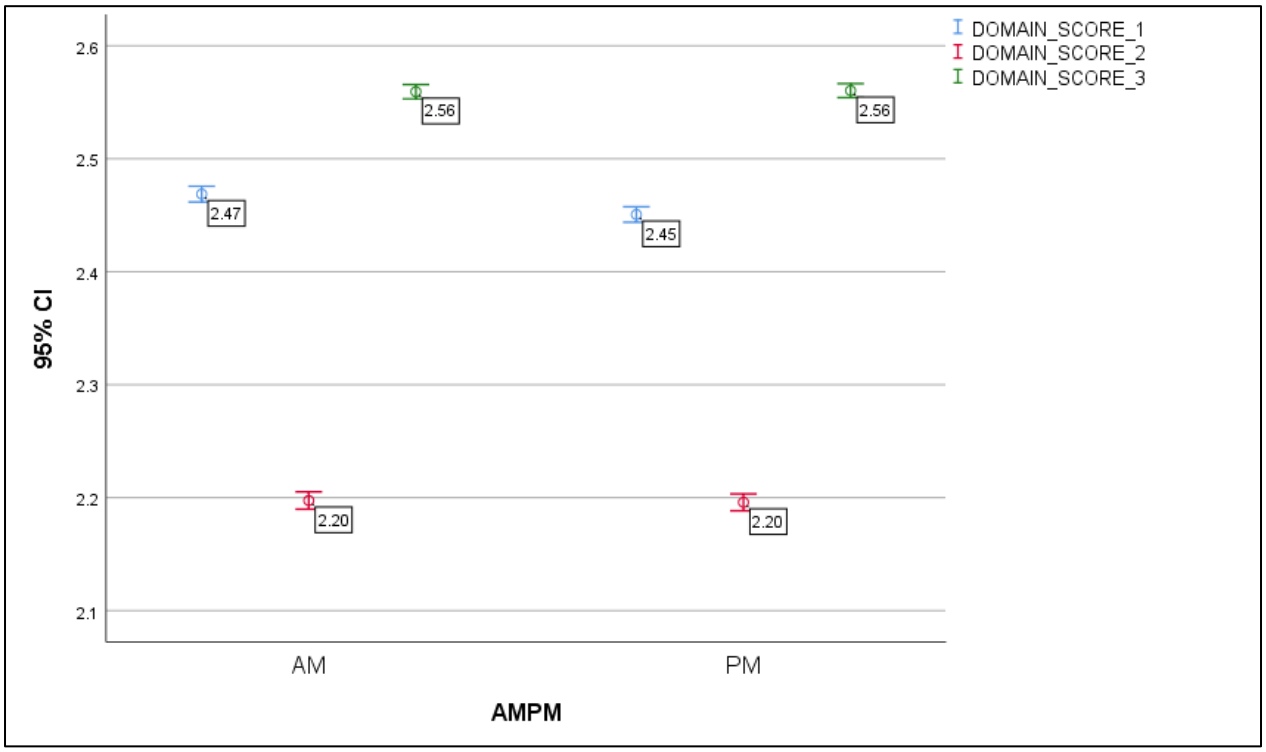

Figure 27: Mean domains scores across AM and PM

[6] Note – this is a simple comparison of means, and the lengths of the error bars is not entirely accurate as the dependency in the data has not been taken fully into account (i.e. the same students across stations, and same examiners, mostly, AM vs. PM). The error bars would, in theory, be shorter than those presented if these dependencies were properly taken into account.

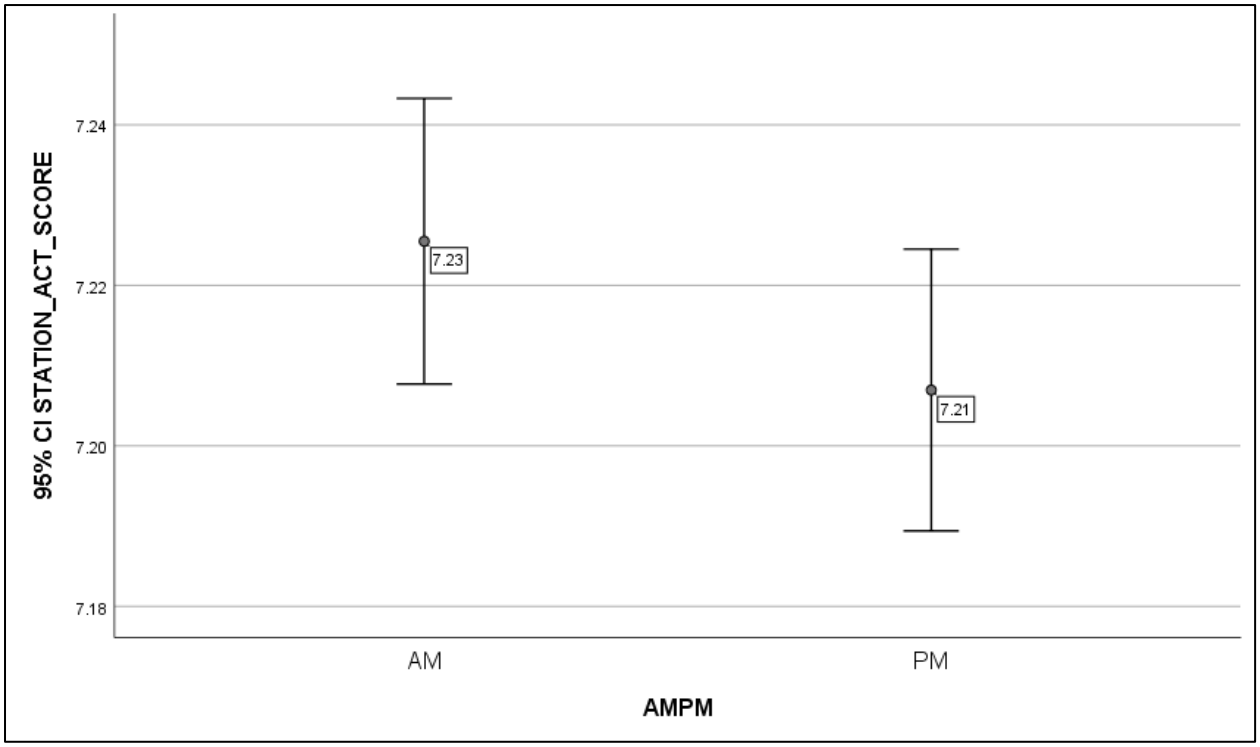

Figure 28: Mean station total score across AM and PM

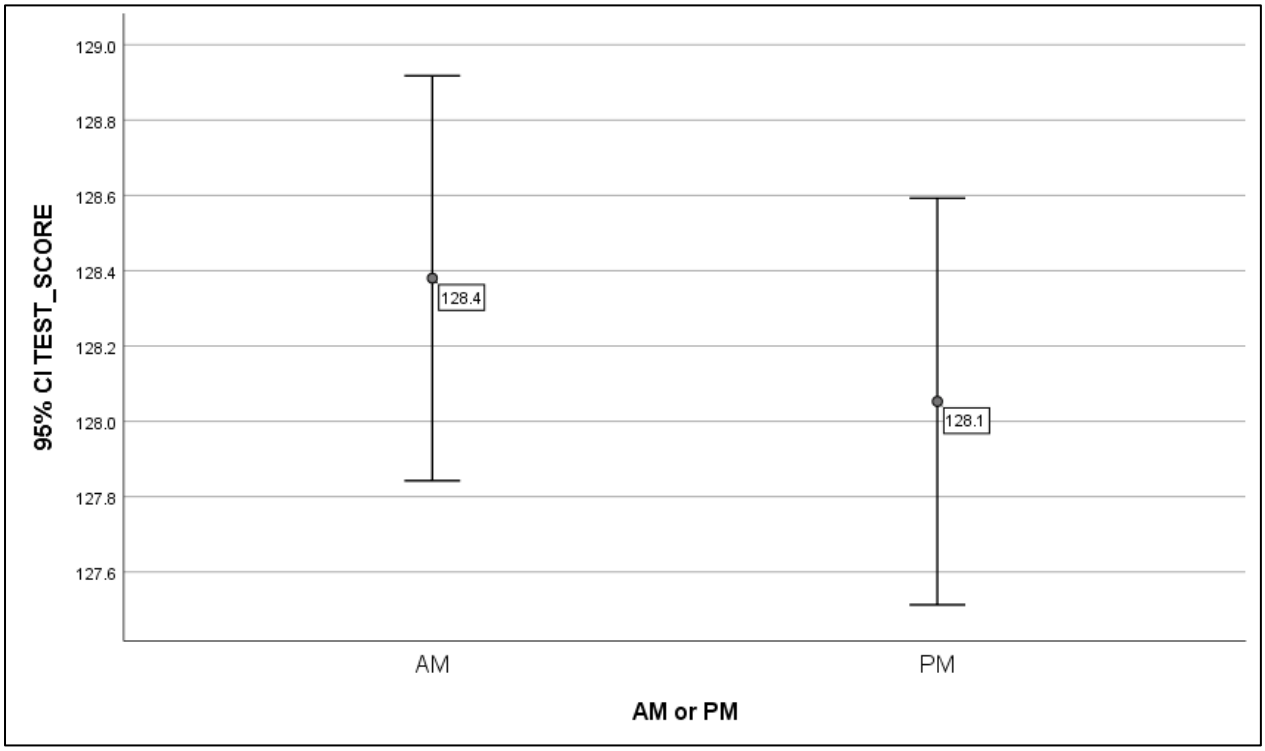

Figure 29: Mean total exam score across AM and PM

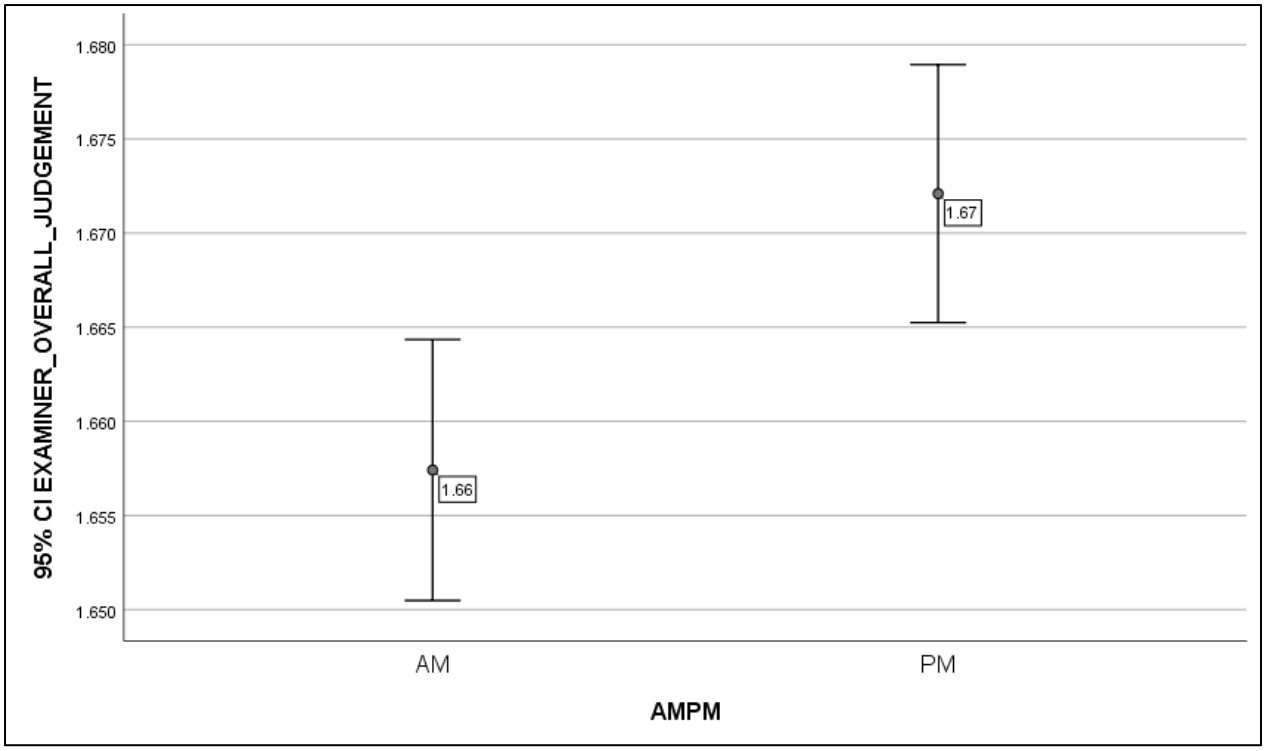

Figure 30: Mean examiner grade across AM and PM

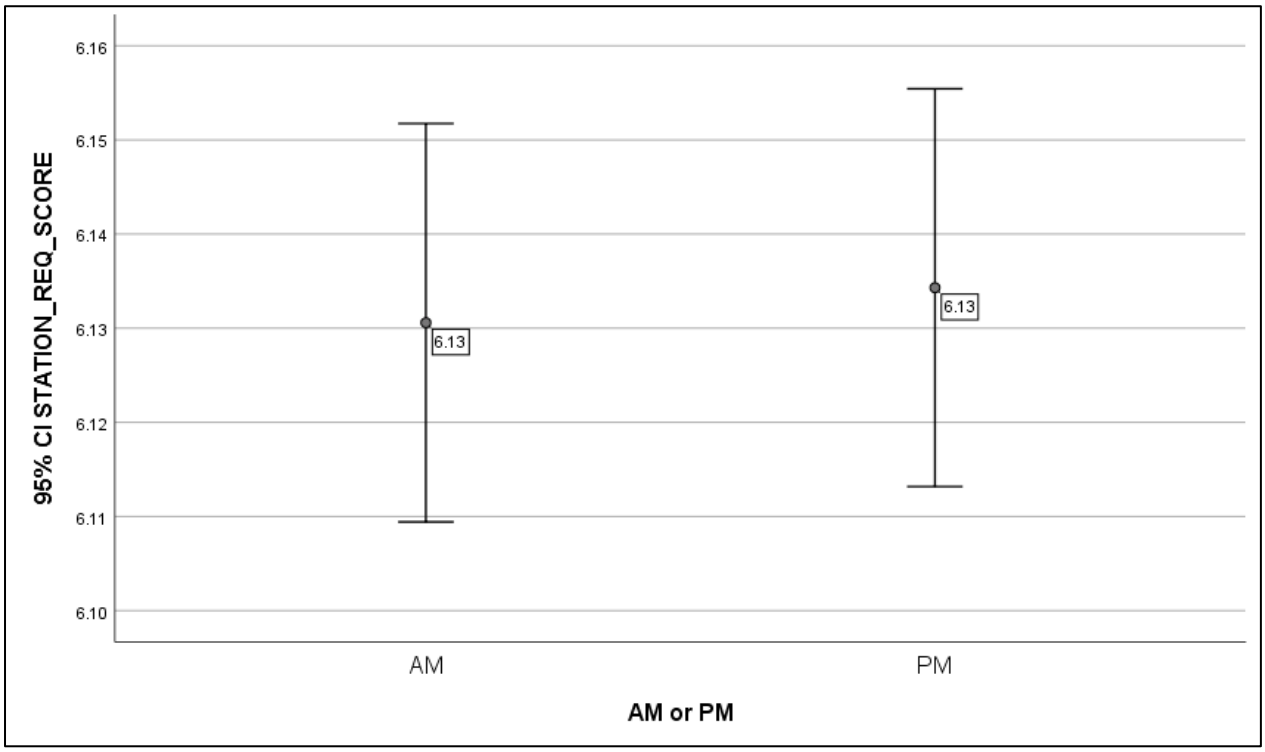

Figure 31: Mean station cut-score across AM and PM

Calculating an exact p-value for each of these differences is quite complex because of the different dependencies in the data. However, a simple calculation of Cohen's d for each of these differences shows that they are very small (all Cohen's $d < 0.02$).

One could imagine doing more complex analysis – e.g. comparing mean examiner scores morning and afternoon for the same examiner, but this is somewhat problematic given that stations and students would be different.

Summary of findings for RQ5

Overall, the comparison of morning and afternoon scoring does not suggest that there is any obvious pattern of difference between the two. That the mean scores in Figure 27 to Figure 31 are so similar gives some re-assurance that examiners are not behaving systematically different across the morning and afternoon sessions. For example, there is no particular evidence of examiner fatigue impacting adversely on the afternoon scoring.

Brief conclusion

Overall, the analysis presented in this report indicates that, on the whole, the PLAB assessments are working well, and there are no obvious serious issues or important recommendations that need to be made in light of the analysis for the five research questions investigated.

The only specific action that might be considered would be to consider reviewing the Angoff judgment process in PLAB1 as highlighted in the summary of findings for the first research question.

References

Clauser, B.E., Mee, J., Baldwin, S.G., Margolis, M.J. and Dillon, G.F. 2009. Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement*. **46**(4), pp.390–407.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Dong, T., Saguil, A., Artino, A.R., Gilliland, W.R., Waechter, D.M., Lopreaito, J., Flanagan, A. and Durning, S.J. 2012. Relationship between OSCE scores and other typical medical school performance indicators: a 5-year cohort study. *Military Medicine*. **177**(9 Suppl), pp.44–46.

Homer, M., Darling, J. and Pell, G. 2012. Psychometric characteristics of integrated multi-specialty examinations: Ebel ratings and unidimensionality. *Assessment & Evaluation in Higher Education*. **37**(7), pp.787–804.

Homer, M., Fuller, R., Hallam, J. and Pell, G. 2019. Setting defensible standards in small cohort OSCEs: Understanding better when borderline regression can 'work'. *Medical Teacher*. **0**(0), pp.1–10.

Park, W.B., Kang, S.H., Lee, Y.-S. and Myung, S.J. 2015. Does Objective Structured Clinical Examinations Score Reflect the Clinical Reasoning Ability of Medical Students? *The American Journal of the Medical Sciences*. **350**(1), pp.64–67.

Pell, G., Fuller, R., Homer, M.S. and Roberts, T. 2010. How to measure the quality of the OSCE: A review of metrics. *Medical Teacher*. **32**(10), pp.802–811.

Trafimow, D. 2016. The attenuation of correlation coefficients: a statistical literacy issue. *Teaching Statistics*. **38**(1), pp.25–28.

Wasserstein, R.L. and Lazar, N.A. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*. **70**(2), pp.129–133.