Prescribing Safety Assessment 2022: Psychometric Report

Overview

- This report analyses data from 8,477 undergraduate sittings of the 2022 PSA exam papers in UK medical schools.
- The data are from UK medical schools (undergraduates) only.
- The PSA was administered on four dates: 4th February; 14th March; 29th April; 6th June.
- Four papers were used (with a fifth paper reserved for remote use by students who were isolating).
- Schools were allocated papers based on previous performance of previous cohorts on the PSA and based on whether students were sitting in the morning and/or afternoon sessions.
- Standards were set by an *a priori* Angoff process and were not modified post-hoc.

Executive Summary

- Angoff standards were slightly higher in the four PSA papers than in previous years, suggesting that this year's papers may be slightly easier.
- Overall performance on the four test papers administered was slightly higher than in previous years reinforcing the difference in Angoff standards set.
- Pass rates at first sittings were also higher than in previous years – pass rates ranging from 95.6% (paper A) to 92.6% (paper D). Failure rates were below 10% for all four papers.
- Analysis of all first sittings data showed that scores on Paper A were higher than on the other three papers. Scores on papers B, C and D were broadly equivalent, with scores on paper C being marginally the lowest. This is in line with the Angoff standard setting.
- The four test papers showed reasonably good reliability for a 60 item test, with Cronbach's alpha for internal consistency at or very near 0.8.
- No items were answered correctly or incorrectly by all students. All items had good item-level statistics.
- Item analyses did not highlight any items which stand out as obviously negatively impacting the assessments to any significant degree. A few items were flagged for review on discrimination statistics.
- Analyses of the passing standards against reliability measures, overall cohort performance and test equating using the anchor items following the February sittings suggested that the Angoff standards for the four papers were appropriate. These analyses were repeated for all sittings following the June sittings with very similar results.
- There was some evidence of differential attainment across schools. Failure rates at first sittings ranged from 0.4% to 36.8%. Two schools had first sitting failure rates over 25%.
- As in previous years, resitting students scored significantly lower than first sitting students.

Recommendations

- There was no evidence following psychometric analyses of the February sitting data to suggest that any items should be removed from any papers. Some items with very low facility and/or negative discrimination statistics were highlighted for review, which was completed after the February sitting.
- Psychometric analysis of item data from all sittings did not highlight any new items for review.

1. Summary Assessment Analysis

The data in this report are from 8,477 sittings (8,078 first sittings and 399 resits) of the PSA in 2022 involving students from 34 UK medical schools. Standards were set by an Angoff process in advance of the first sitting, and were not modified post-hoc.

Four test papers (papers A, B, C and D) were used across all four test dates and were delivered to students under exam conditions in medical schools with face-to-face invigilation. A fifth paper (paper E) was reserved for remote sittings for students in isolation but was only required for use by seven students over the year.

Schools were ranked by PSA performance quintile from the past three years and assigned to the four papers to approximate an equal distribution of schools by historical performance across all four papers. Where schools had split cohorts sitting across morning and afternoon sessions, different papers were delivered in the morning and afternoon.

1.1 Student Numbers by Sitting Date

Appendix 1 details the number of sittings (first sittings and resits) for each of the 34 UK medical schools – and also shows the anonymised school code for reference in the rest of this report.

	February	March	April	June	Total
First Sitting					
Paper A	1025	660	598	2	2285
Paper B	1494	791	25	29	2339
Paper C	1095	525	0	0	1620
Paper D	1428	386	7	6	1827
Paper E	1	5	1	0	7
Resits					
Paper A		38	104	20	162
Paper B		2	60	102	72
Paper C		0	0	4	4
Paper D		9	15	45	69
Paper E		0	0	0	0
Total Sittings					
First Sittings	5043	2367	631	37	8078
Resits	0	49	179	171	399
All Sittings	**5043**	**2416**	**810**	**208**	**8477**

1.2 PSA Summary Statistics

2022 Key Statistics

The following table summarises the overall aggregated performance statistics for each paper across all four event dates, including a breakdown of first sittings and resits.

	First Sittings				Resits				All Sittings			
	Paper A	**Paper B**	**Paper C**	**Paper D**	**Paper A**	**Paper B**	**Paper C**	**Paper D**	**Paper A**	**Paper B**	**Paper C**	**Paper D**
Descriptives												
Students (N)	2285	2339	1620	1827	162	164	4	69	2447	2503	1624	1896
Mean Score (%)	163.6 (81.8%)	159.3 (79.7%)	157.5 (78.8%)	159.3 (79.6%)	146.8 (73.4%)	143 (71.5%)	146.5 (73.2%)	130.7 (65.3%)	162.4 (81.2%)	158.3 (79.1%)	157.5 (78.8%)	158.2 (79.1%)
Median Score (%)	166 (83%)	162 (81%)	160 (80%)	162 (81%)	147 (73.5%)	145 (72.5%)	143 (71.5%)	130 (65%)	165 (82.5%)	161 (80.5%)	160 (80%)	162 (81%)
Std Deviation (%)	17.1 (8.6%)	18.2 (9.1%)	18.4 (9.2%)	20.6 (10.3%)	18 (9%)	17.1 (8.5%)	13.8 (6.9%)	21.8 (10.9%)	17.7 (8.8%)	18.6 (9.3%)	18.4 (9.2%)	21.3 (10.6%)
Minimum Score (%)	79 (39.5%)	63 (31.5%)	38 (19%)	61 (30.5%)	104 (52%)	83 (41.5%)	135 (67.5%)	85 (42.5%)	79 (39.5%)	63 (31.5%)	38 (19%)	61 (30.5%)
Maximum Score (%)	199 (99.5%)	196 (98%)	195 (97.5%)	197 (98.5%)	185 (92.5%)	180 (90%)	165 (82.5%)	169 (84.5%)	199 (99.5%)	196 (98%)	195 (97.5%)	197 (98.5%)
Pass Rate	95.60%	95.10%	94.80%	92.60%	82.10%	82.90%	100%	56.50%	94.70%	94.30%	94.80%	91.20%
Reliability												
Pass Mark	65.50%	63.50%	63%	64%	65.50%	63.50%	63%	64%	65.50%	63.50%	63%	64%
Cronbach's Alpha	0.794	0.783	0.805	0.814	0.745	0.714	na	0.742	0.804	0.79	0.804	0.823
Standard Error of Measurement (%)	3.88%	4.24%	4.07%	4.44%	4.55%	4.57%	na	5.53%	3.91%	4.25%	4.07%	4.47%

The following three tables show key summary statistics from 2021, 2020 and 2019 PSA papers for comparison purposes. Passing scores, pass rates, reliability measures and overall performance in both first sittings and resits were comparable in the four 2022 PSA papers to papers administered in previous years.

2021 Key Statistics

	First Sittings			Resits			All Sittings		
	Paper A Face-to-face	**Paper B Remote - proctored**	**Paper C Remote - unsupervised**	**Paper A Face-to-face**	**Paper B Remote - proctored**	**Paper C Remote - unsupervised**	**Paper A Face-to-face**	**Paper B Remote - proctored**	**Paper C Remote - unsupervised**
Descriptives									
Students (N)	1572	2897	3110	165	192	94	1737	3089	3204
Mean Score (%)	156.1 (78.1%)	151.4 (75.7%)	153.3 (76.6%)	144.5 (72.3%)	136.7 (68.3%)	136.6 (68.3%)	155 (77.5%)	150.5 (75.2%)	152.8 (76.4%)
Median Score (%)	159 (79.5%)	154 (77.0%)	155 (77.5%)	146 (73.0%)	137 (68.5%)	138 (69.0%)	158 (79.0%)	153 (76.5%)	155 (77.5%)
Standard Deviation (%)	18.2 (9.1%)	20.3 (10.1%)	16.6 (8.3%)	18 (9.0%)	19.3 (9.7%)	18.2 (9.1%)	18.5 (9.3%)	20.5 (10.3%)	16.9 (8.5%)
Minimum Score (%)	43 (21.5%)	16 (8.0%)	74 (37.0%)	82 (41.0%)	55 (27.5%)	93 (46.5%)	43 (21.5%)	16 (8.0%)	74 (37.0%)
Maximum Score (%)	194 (97.0%)	197 (98.5%)	191 (95.5%)	184 (92.0%)	181 (90.5%)	176 (88.0%)	194 (97.0%)	197 (98.5%)	191 (95.5%)
Pass Rate	93.3%	92.9%	93.6%	87.3%	79.7%	66.0%	92.7%	92.1%	92.8%
Reliability									
Pass Mark	62.5%	60.5%	63.5%	62.5%	60.5%	63.5%	62.5%	60.5%	63.5%
Cronbach's Alpha	0.765	0.806	0.744	0.68	0.735	0.723	0.766	0.807	0.751
Standard Error of Measurement (%)	4.42%	4.47%	4.20%	5.10%	4.98%	4.79%	4.48%	4.51%	4.22%

2020 key statistics

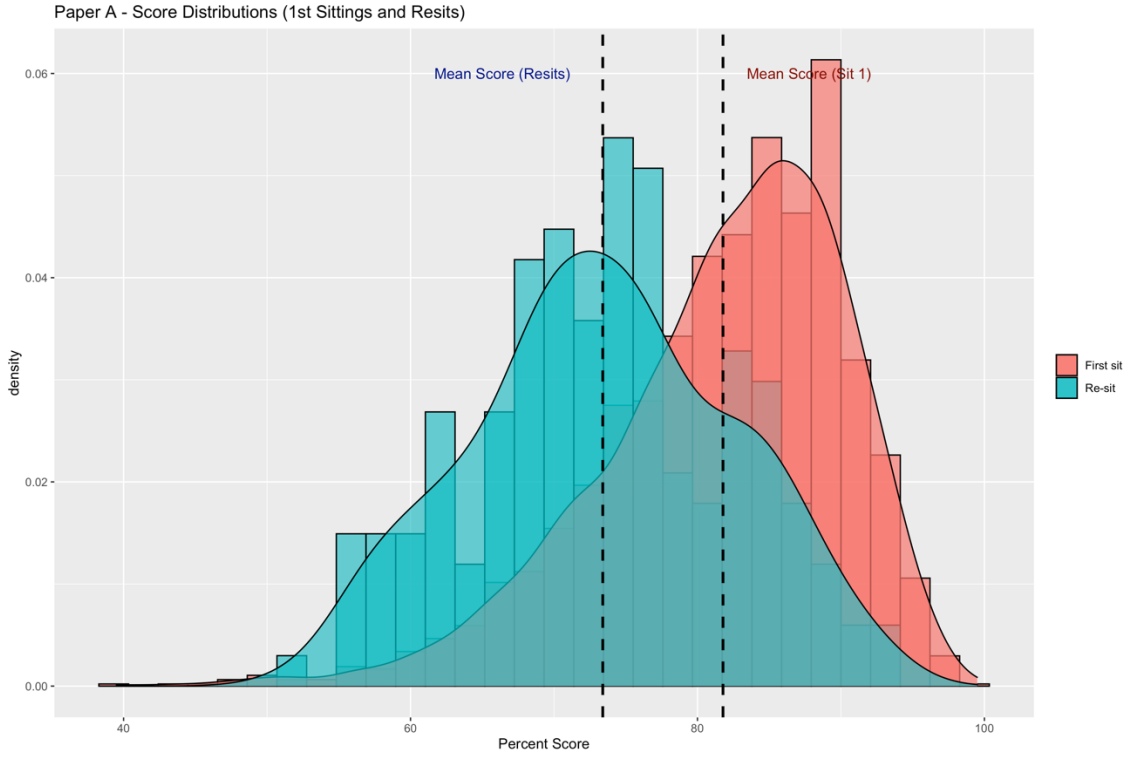
Cohort		Paper A	Paper B	Paper C	Paper D	Paper F	Paper S
	Pass mark	61%	57%	60%	62%	62%	63%
	Cronbach's Alpha	0.8	0.78	0.77	0.79	0.74	0.72
	Standard Error	6.64	6.28	6.46	6.34	6.4	5.58
All	Number of students	1488	1460	1571	1529	1484	560
	Mean (SD)	150 (20.4)	152 (19.0)	140 (19.0)	151 (19.5)	152 (17.0)	163 (15.2)
	Pass Rate	91%	97%	86%	91%	93%	98%
First sit	Number of students	1420	1441	1560	1518	1214	453
	Mean (SD)	151 (20.3)	152 (18.8)	140 (19.4)	151 (19.4)	155 (17.0)	165 (13.6)
	Pass Rate	92%	97%	86%	91%	95%	99%
Resit	Number of students	68	19	11	11	270	107
	Mean (SD)	135 (15.7)	137 (15.6)	132 (14.9)	125 (15.9)	143 (17.6)	151 (16.1)
	Pass Rate	79%	95%	82%	36%	85%	93%

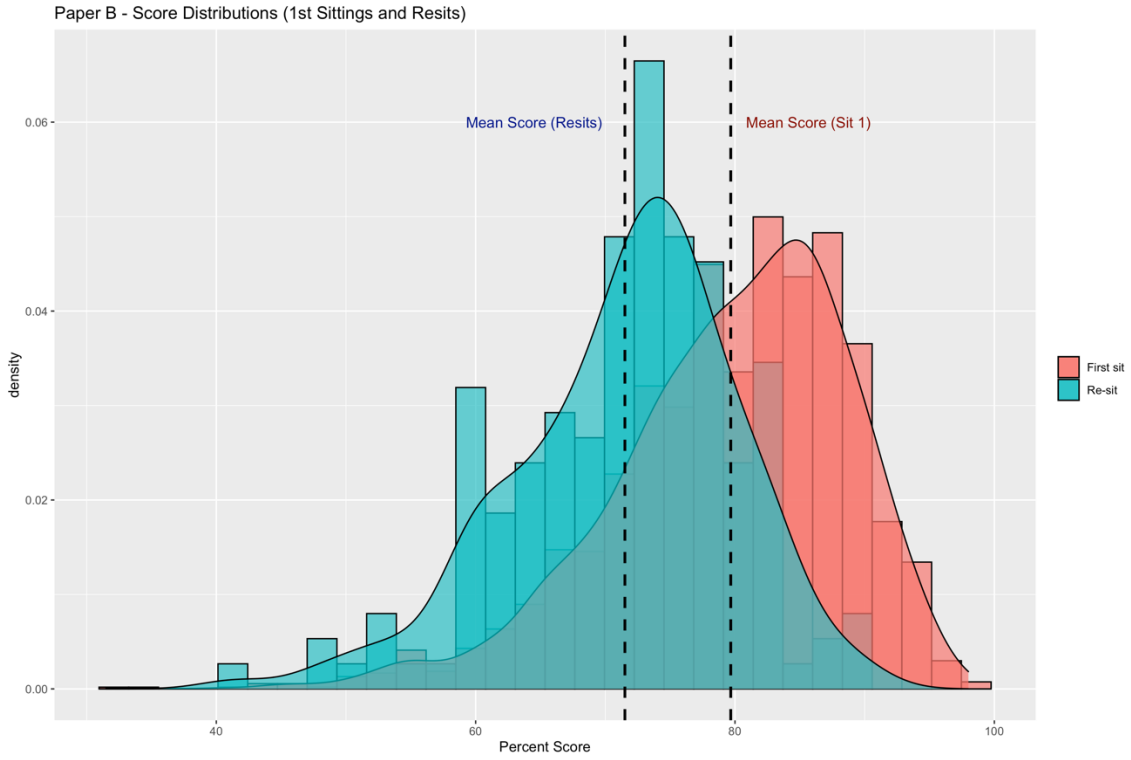
2019 key statistics

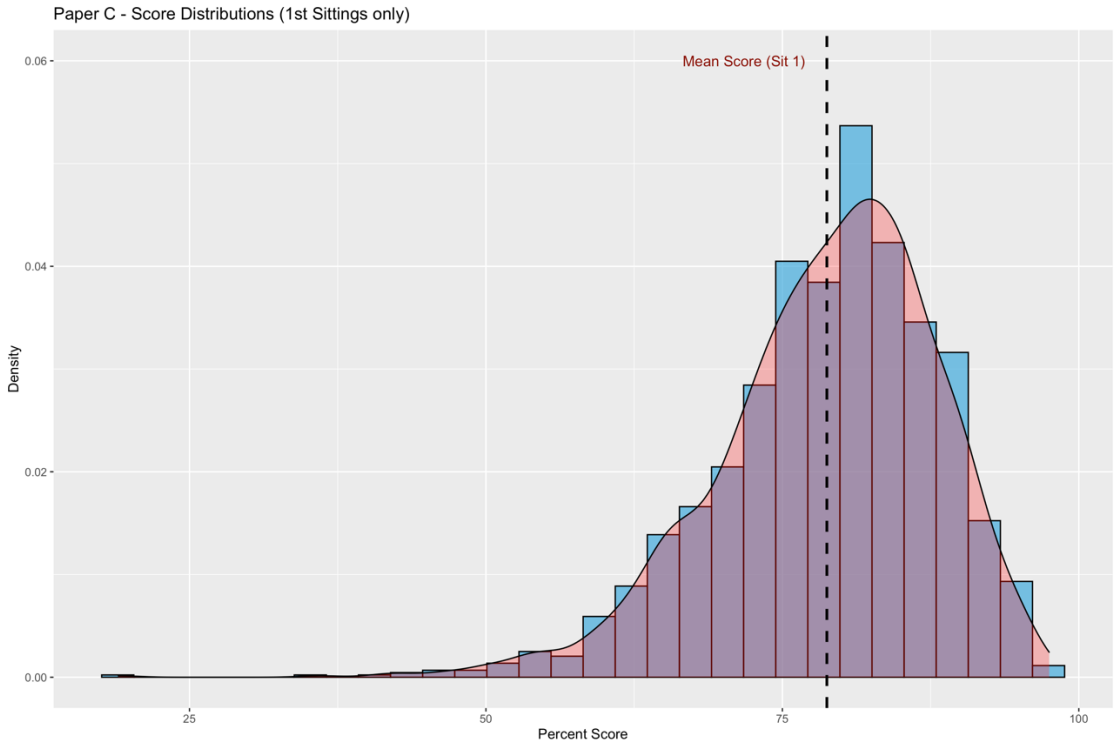
Cohort		Paper A	Paper B	Paper C	Paper D
	Pass mark	63%	62%	62.5%	62.5%
	Cronbach's Alpha	0.8	0.8	0.81	0.83
All	Number of students	2354	1932	1800	1734
	Mean (SD)	159.6 (17.7)	162.7 (17.2)	155.1 (18.9)	161.8 (17.7)
	Pass Rate	95.7%	97.4%	92.3%	96.4%
First sit	Number of students	2293	1816	1755	1660
	Mean (SD)	160.3 (16.0)	164.0 (16.2)	155.8 (18.3)	163.0 (16.5)
	Pass Rate	96.6%	98.5%	93.3%	97.7%
Resit	Number of students	61	116	45	74
	Mean (SD)	133.2 (18.7)	141.3 (18.8)	127.0 (20.5)	133.8 (19.6)
	Pass Rate	62.3%	80.2%	51.1%	66.2%

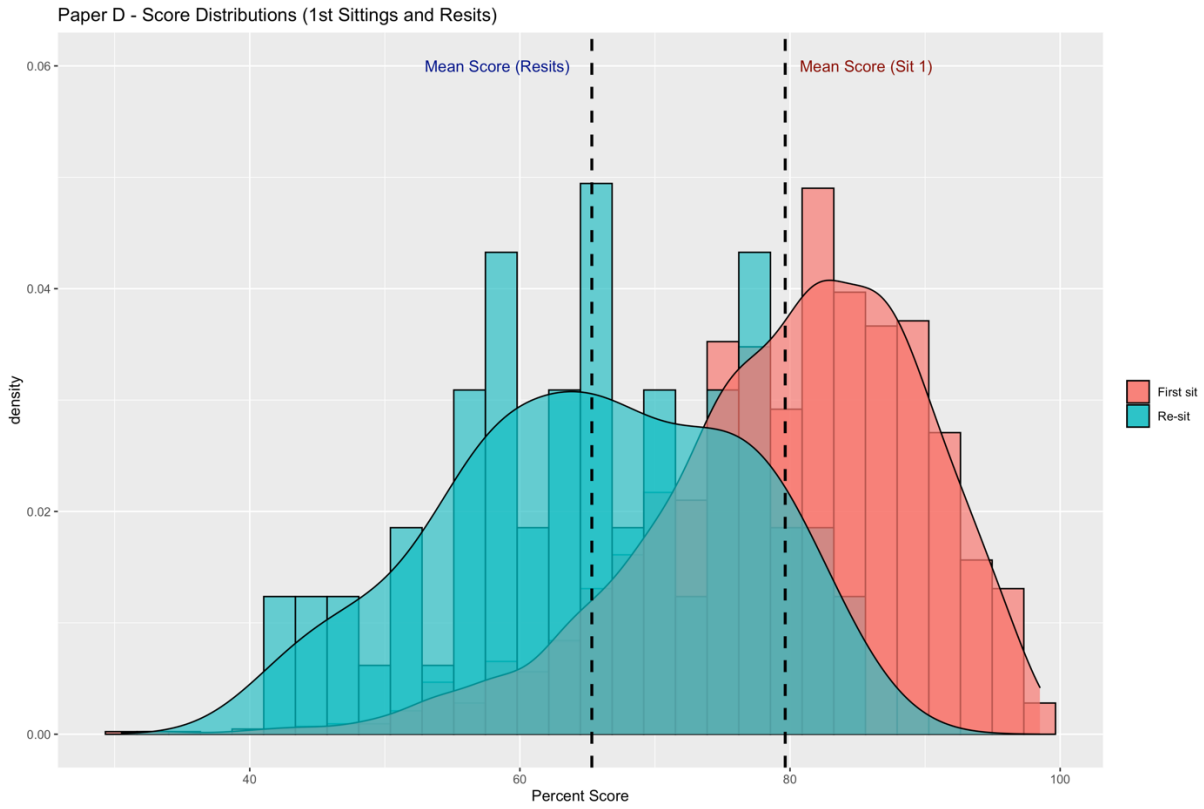
1.3 Total Score Distributions

The histograms below show the percentage score distributions for each paper (first sittings and resits – all test dates)- the dotted vertical lines indicating the group mean scores on each paper. First sitting scores were negatively skewed on all three PSA papers. Resit scores are significantly lower and more normally distributed around the group mean on all papers. Note that paper C was only used for four resit events and this data is excluded accordingly.

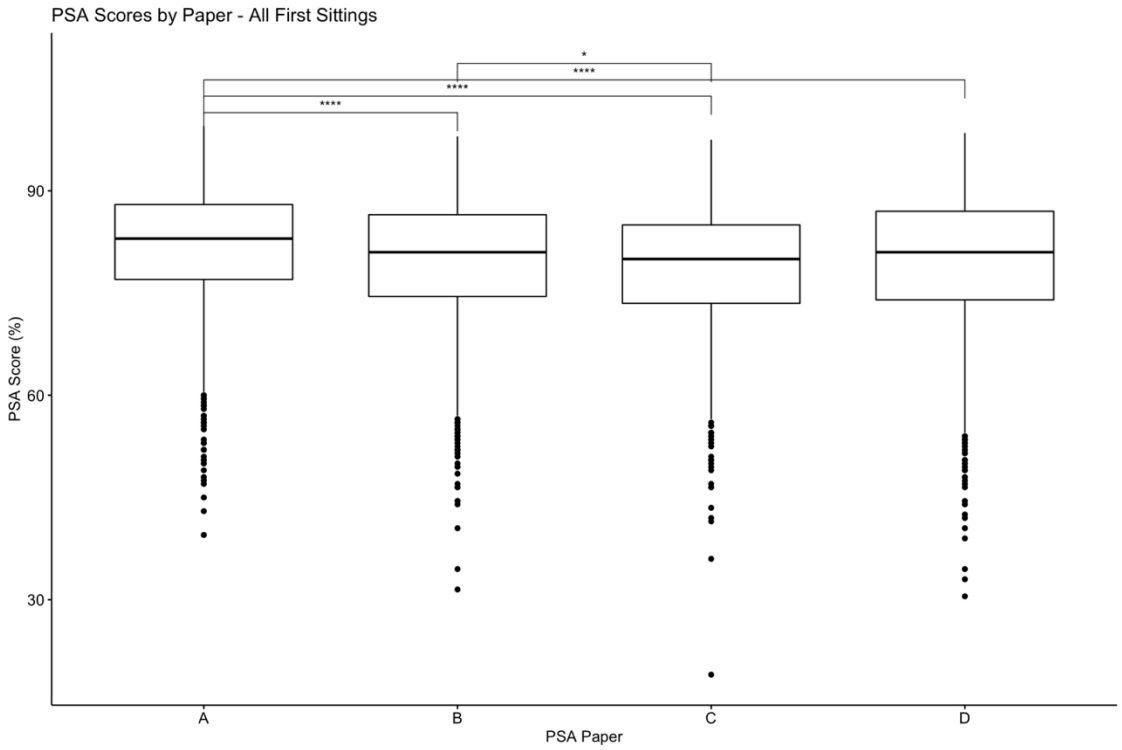




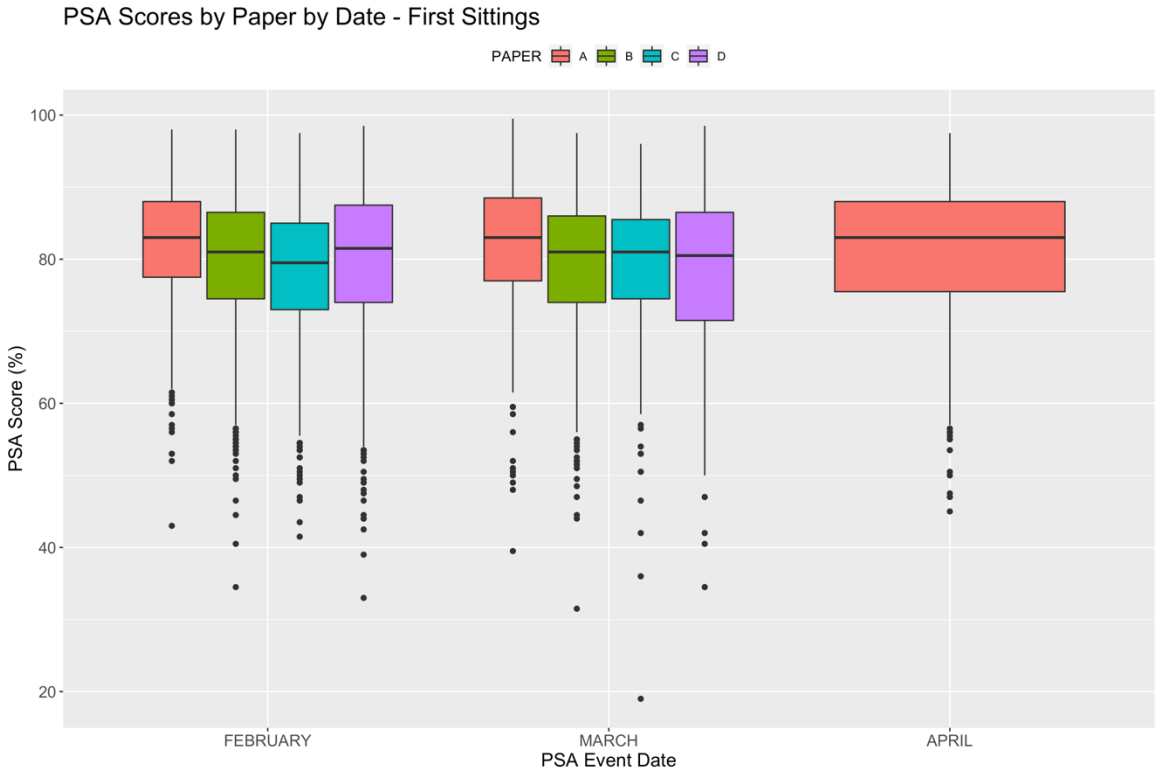




The following plot shows the distribution of all first sitting scores in each test paper. Independently scores on paper A were significantly higher than the other three papers across all first sittings. Scores on paper C were lower than scores on paper B overall, but otherwise scores on papers B, C and D were not significantly different.



The following plot shows the distribution of first sittings test scores by paper at each sitting date. Data from sittings where papers were sat by fewer than 30 candidates have been removed from this plot. It can be seen again here that paper A had higher scores than the other three papers at the two main first sitting dates (February and March). Scores on paper C were lower at the February sitting than papers B and C, but at the March sitting there was no significant difference in scores between papers B, C and D.



1.4 Performance by PSA Section

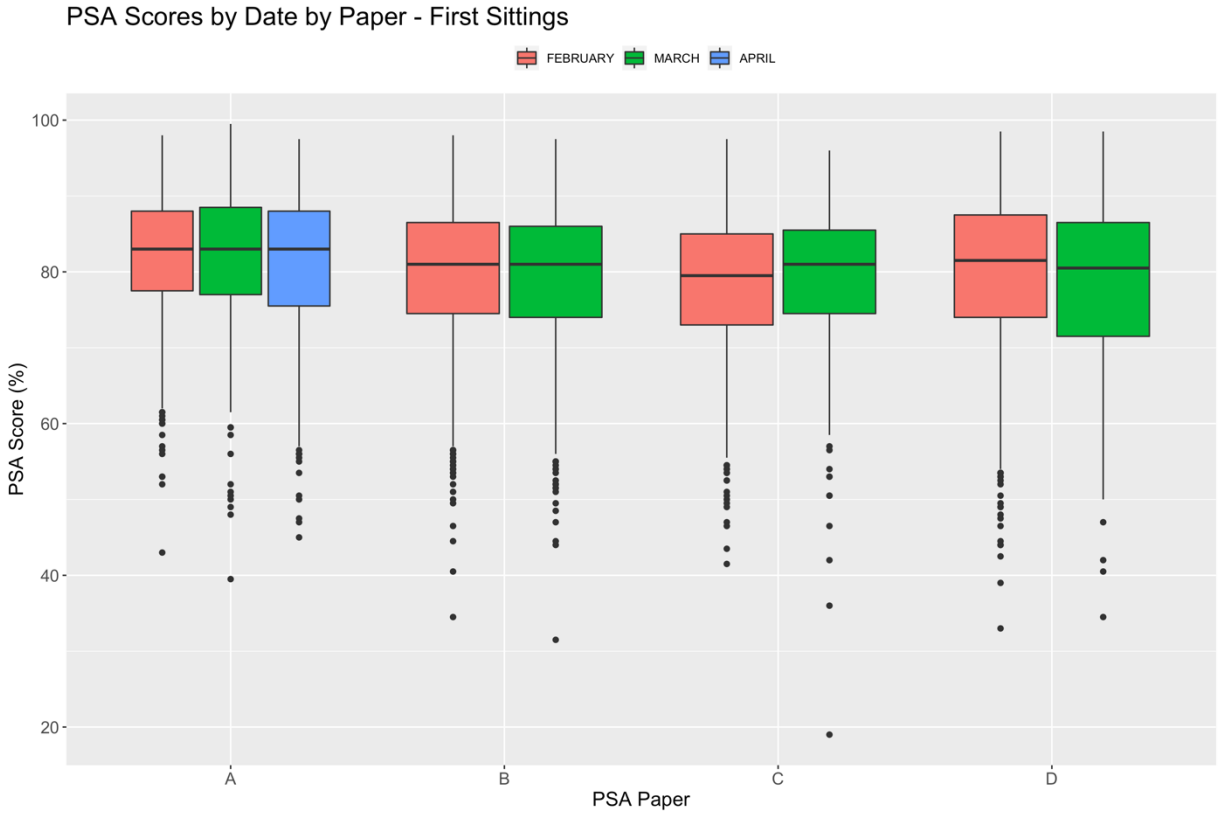
The PSA contains eight distinct sections across a range of clinical contexts: prescribing (PWS); prescription review (REV); planning management (MAN); providing information about medicines (COM); calculation skills (CAL); adverse drug reactions (ADR); drug monitoring (TDM); and data interpretation (DAT). The following table shows average performance on each section of the PSA by paper highlighting sections where students score one standard deviation greater (green) or less (red) than the mean score for the sitting. Note the small numbers of students taking resits should be taken into account and no cell is highlighted for paper C resits as only 4 candidates sat this paper. This indicative analysis suggests that students found the planning management questions particularly challenging in all papers – and this is in line with findings in previous years. Students scored higher than the mean at first sittings on all papers on the prescription review items and on the drug monitoring items. There was some variation in performance across section by paper but less so than in previous years suggesting that the papers were generally well balanced – however students scored higher on the 10 point prescribing items on Paper A (10% points higher than on Paper C) – which may go some way to account for the overall observation that this was the easiest of the papers.

	First Sittings				Resits				All Sittings			
	Paper A	**Paper B**	**Paper C**	**Paper D**	**Paper A**	**Paper B**	**Paper C**	**Paper D**	**Paper A**	**Paper B**	**Paper C**	**Paper D**
PSA Paper Section												
PWS Avg % Score (SD)	85.6 (10.9)	84 (12.3)	76.5 (12.2)	79.4 (14.5)	80.6 (12.2)	77.6 (12.5)	*66.2 (6.2)*	65.2 (16.9)	85.3 (11)	83.6 (12.4)	76.5 (12.2)	78.9 (14.8)
REV Avg % Score (SD)	82.3 (10.4)	80.9 (8.8)	84.8 (9.5)	82.8 (9.9)	74.7 (11.2)	76 (8.4)	*82.8 (8.3)*	71.3 (10.2)	81.8 (10.6)	80.6 (8.8)	84.8 (9.5)	82.3 (10.1)
MAN Avg % Score (SD)	68.4 (16.6)	65.1 (17)	77.7 (16.8)	77.5 (16.9)	58.2 (18.2)	55.6 (15.8)	*59.4 (12)*	62.9 (15.9)	67.7 (16.9)	64.5 (17.1)	77.7 (16.8)	77 (17.1)
COM Avg % Score (SD)	84.9 (15.9)	71.7 (18.4)	76.2 (17.9)	81 (18.4)	72.4 (19.2)	62.2 (20)	*66.7 (13.6)*	62.1 (21)	84.1 (16.5)	71.1 (18.6)	76.2 (17.9)	80.3 (18.8)
CAL Avg % Score (SD)	77.5 (20)	76.8 (20.6)	75 (22.5)	74.9 (21.1)	59.6 (25.6)	61.8 (26.5)	*68.8 (26)*	54.7 (27.4)	76.3 (20.9)	75.8 (21.3)	75 (22.5)	74.2 (21.7)
ADR Avg % Score (SD)	81.4 (13.5)	76.9 (16.9)	83.4 (14.5)	77.6 (17.7)	75.2 (14.3)	67.4 (18.1)	*93.8 (12.5)*	68.5 (18.9)	81 (13.6)	76.2 (17.2)	83.5 (14.5)	77.3 (17.8)
TDM Avg % Score (SD)	83.1 (16.4)	83.9 (14.5)	84.1 (15.9)	82.9 (16.3)	73 (20.4)	75.1 (17.2)	*90.6 (12)*	69.4 (18.1)	82.5 (16.9)	83.3 (14.9)	84.1 (15.9)	82.4 (16.6)
DAT Avg % Score (SD)	73.8 (19.4)	76.6 (20.3)	73.6 (20.4)	79.2 (20.1)	59.8 (22.3)	62.8 (22.2)	*75 (16.7)*	61.1 (23)	72.9 (19.9)	75.7 (20.7)	73.6 (20.4)	78.5 (20.5)

1.5 Summary Statistics by Paper by Sitting

This section considers whether there are any significant differences in the score distributions of each paper over the year. Significant growth in score distributions on individual papers across the year may be indicative of question paper leakage, but variations in paper performance across time may be explained by other factors, such as variations in attainment between schools and whether schools use the PSA summatively or formatively, for example.

The boxplots below show the distribution of first sitting scores for each paper at each date (excluding sittings with fewer than 30 students). Scores on papers A and B did not change over time. Scores on paper C increased in the March sitting compared to the February sitting whereas scores on paper D decreased in the March sitting compared to the February, but in multivariable regression analyses adjusting for school, there were no significant differences in scores over time on any paper.



The following tables outline summary statistics by sitting date for each paper (first sittings only).

Paper A First Sittings	February	March	April	June
Num. Students	1025	660	598	2
Mean Score (%)	82.2%	81.9%	80.9%	82.2%
SD (%)	7.8%	8.7%	9.6%	11.7%
Pass Rate	97.1%	97.0%	91.6%	100%

Paper B First Sittings	February	March	April	June
Num. Students	1494	791	25	29
Mean Score (%)	79.8%	79.5%	77.0%	77.9%
SD (%)	8.8%	9.4%	10.6%	10.6%
Pass Rate	95.6%	94.4%	88.0%	89.7%

Paper C First Sittings	February	March	April	June
Num. Students	1095	525	0	0
Mean Score (%)	78.50%	79.40%		
SD (%)	9.10%	9.30%		
Pass Rate	94.10%	96.40%		

Paper D First Sittings	February	March	April	June
Num. Students	1428	386	7	6
Mean Score (%)	80.10%	78.40%	63.60%	76.30%
SD (%)	9.90%	10.90%	20.90%	13.80%
Pass Rate	93.80%	88.90%	42.90%	83.30%

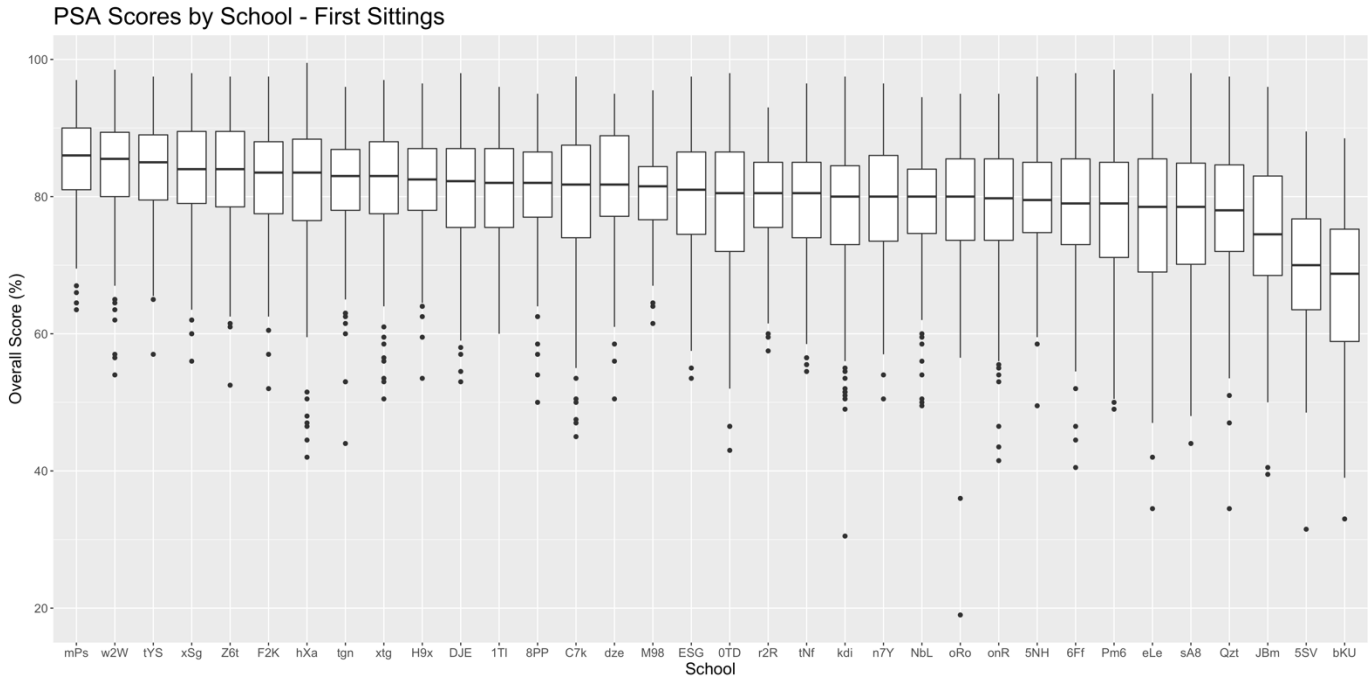
3. Overall Performance by School

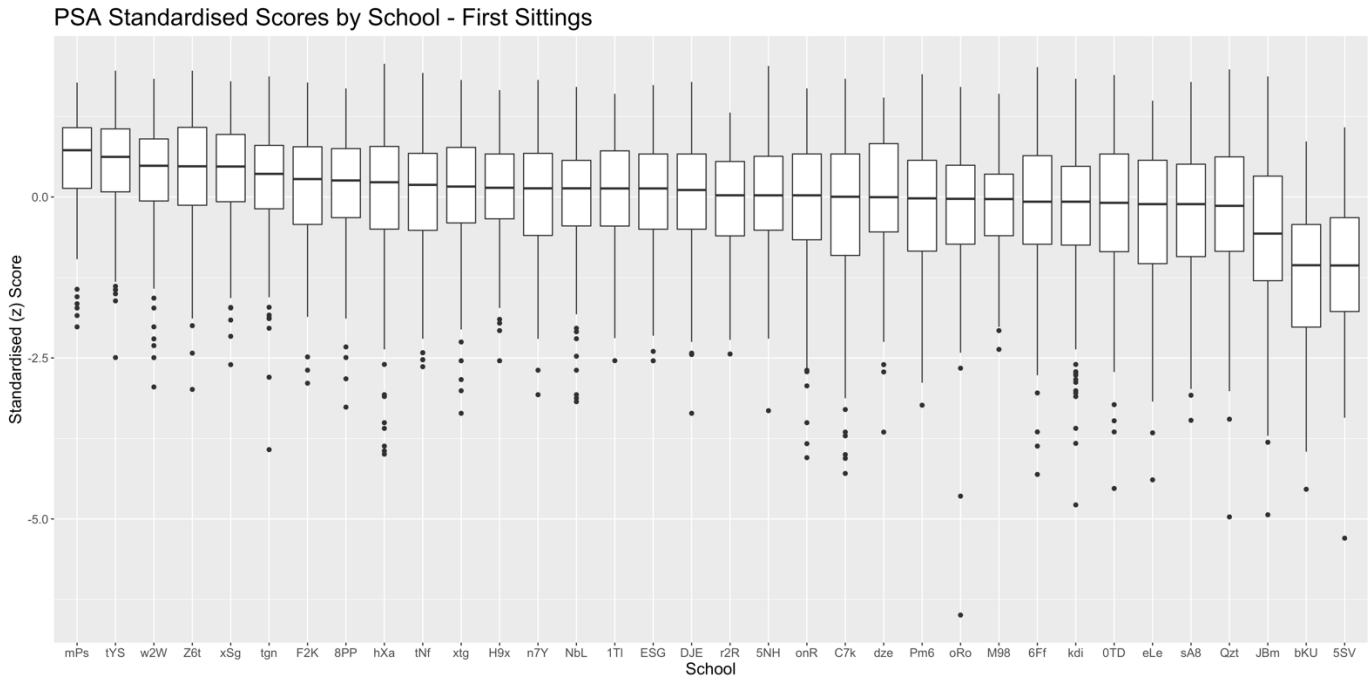
34 UK medical schools took part in the 2022 PSA sittings. Appendix 1 details the number of sittings and papers used (first sittings and resits) for each school. The school associated with the anonymised 3-digit school code used in this section is also shown in Appendix 1.

3.1 Scores by School

The table below shows mean and median scores and failure rates at first sittings for all schools (by school code). The plots on the next page show the variation in overall first sitting test scores across schools. Plots are ordered left to right by overall school performance (highest to lowest median scores). The first plot shows overall PSA score distributions and the second plot shows standardised score (z-score) distributions – standardised scores are generally most suitable for comparing performance across schools where different papers are taken by different students. Appendix 4 shows these data further disaggregated by test paper.

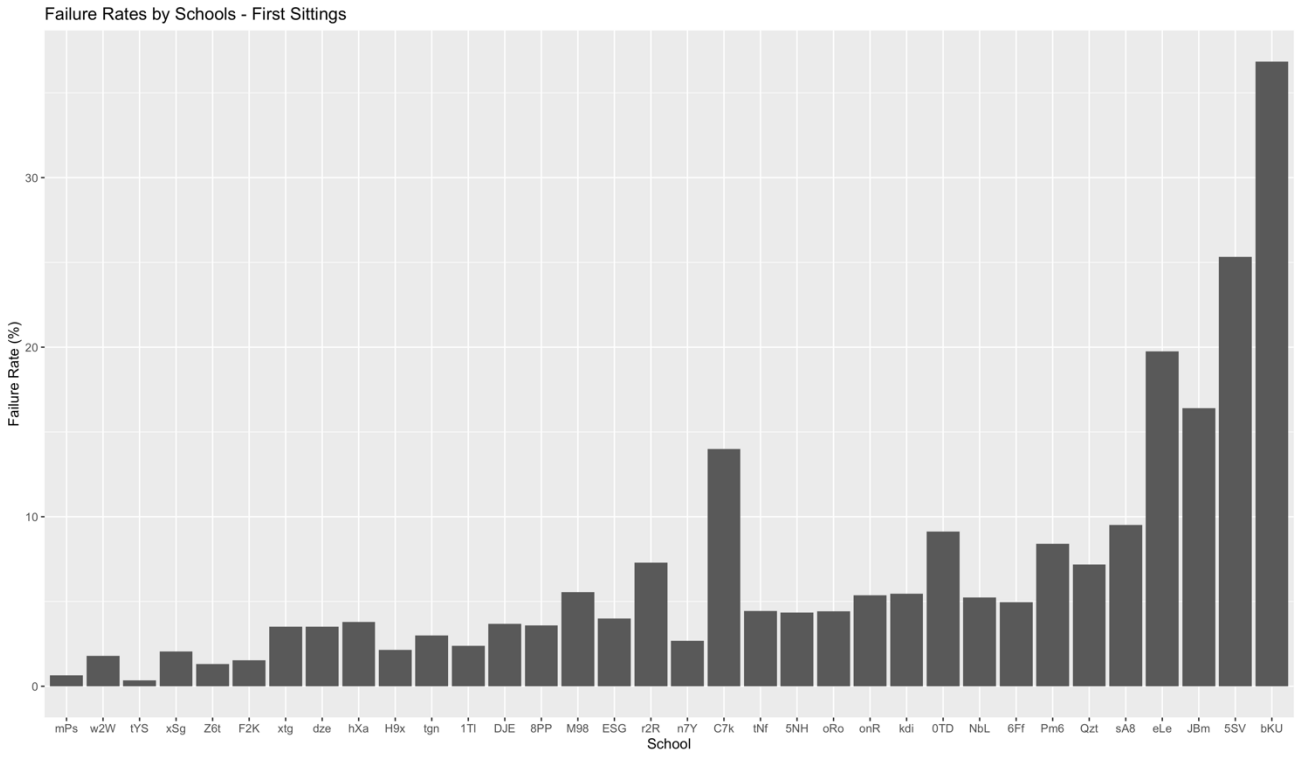
School Code	N Students	Mean Score (%)	Median Score (%)	Median Z-Score	N Fails	Fail Rate (%)
mPs	153	85.1	86	0.726	1	0.7
tYS	274	84	85	0.623	1	0.4
w2W	334	84	85.5	0.484	6	1.8
Z6t	306	83.1	84	0.476	4	1.3
xSg	293	83.3	84	0.473	6	2
tgn	234	81.7	83	0.359	7	3
F2K	326	82.7	83.5	0.278	5	1.5
8PP	139	80.5	82	0.256	5	3.6
hXa	422	81.7	83.5	0.229	16	3.8
tNf	270	79.4	80.5	0.188	12	4.4
xtg	256	81.9	83	0.162	9	3.5
H9x	187	81.7	82.5	0.143	4	2.1
NbL	210	78.7	80	0.134	11	5.2
n7Y	297	79.5	80	0.134	8	2.7
ESG	276	80	81	0.132	11	4
1Tl	209	81.4	82	0.132	5	2.4
DJE	190	81	82.2	0.108	7	3.7
r2R	96	79.7	80.5	0.026	7	7.3
5NH	115	79.2	79.5	0.025	5	4.3
onR	410	78.9	79.8	0.025	22	5.4
C7k	286	79.4	81.8	0.003	40	14
dze	142	81.8	81.8	-0.003	5	3.5
Pm6	226	77.6	79	-0.021	19	8.4
oRo	362	79	80	-0.029	16	4.4
M98	54	80.1	81.5	-0.033	3	5.6
6Ff	464	78.6	79	-0.074	23	5
kdi	348	78.7	80	-0.074	19	5.5
0TD	263	78.7	80.5	-0.091	24	9.1
eLe	81	76	78.5	-0.111	16	19.8
sA8	126	77.4	78.5	-0.111	12	9.5
Qzt	376	77.6	78	-0.138	27	7.2
JBm	195	74.7	74.5	-0.569	32	16.4
bKU	76	66.9	68.8	-1.06	28	36.8
5SV	75	69.4	70	-1.064	19	25.3


PSA Scores by School - First Sittings
Overall Score (%)
100
80
60
40
20
mPs w2W tYS xSg Z6t F2K hXa tgn xtg H9x DJE 1TI 8PP C7k dze M98 ESG 0TD r2R tNf kdi n7Y NbL oRo onR 5NH 6Ff Pm6 eLe sA8 Qzt JBm 5SV bKU
School


PSA Standardised Scores by School - First Sittings
Standardised (z) Score
0.0
-2.5
-5.0
mPs tYS w2W Z6t xSg tgn F2K 8PP hXa tNf xtg H9x n7Y NbL 1TI ESG DJE r2R 5NH onR C7k dze Pm6 oRo M98 6Ff kdi 0TD eLe sA8 Qzt JBm bKU 5SV
School

3.2 Failure Rates by School

The following bar plot shows the variation in failure rates between schools. This is first sitting data. Schools are ordered from left to right by overall school performance (highest to lowest mean PSA scores). Failure rates range from 0.4% to 36.8%.



4. Assessment Reliability, Standard Setting and Failure Rates

A test is said to be homogenous or unidimensional if its items measure a single latent trait or construct, e.g. prescribing skill. The fundamental concept of scale or test reliability assumes that the sample of items in the test are homogenous, and if that assumption is violated it may cause underestimation of test reliability. Factor analyses can be useful in estimating the underlying dimensionality of a test. For the four test papers used in the PSA, parallel analyses were conducted to estimate dimensionality. This analysis used data from first sittings and resits from all test points. (Appendix 2 details the results of these analyses).

For all four papers, there was one principle factor (as can be seen by the "elbows" in the scree plots in Appendix 2) onto which most of the items in the test loaded (contributed information to the underlying principle latent trait). This provided some support for the assumption of unidimensionality. However, it should be noted that the factor analyses did not rule out the presence of other factors in the four test papers, and this may impact on the estimation of test reliability that follows.

Cronbach's alpha ('α') is a measure of scale reliability in unidimensional assessments. Alpha estimates the internal consistency, or how closely related the sets of items are in a test – and therefore tells us how well items work together as a set. A high coefficient suggests that candidates tend to respond in similar ways from one item to the next. Values for alpha range from 0 (where there is no correlation between items) to 1 (where all items correlate perfectly with one another). The widely accepted gold-standard alpha for high stakes examinations is 0.8.

The standard error of measurement (SEM) is an estimate of how repeated measures of a person on the same assessment would be distributed around their theoretical "true" score. The SEM is a function of the reliability of the assessment (α) and the standard deviation in scores on the assessment. Broadly speaking there is around a 66% probability that a candidate's "true score" will be within 1 SEM of their observed score, and around a 95% probability that their "true score" will be within 2 SEMs of their observed score. This is obviously particularly relevant for candidates around the cut-score for an assessment.

The table below shows the alpha and SEM for each paper (all first sittings data).

Reliability Measures	Paper A	Paper B	Paper C	Paper D
Cronbach's Alpha	0.79	0.78	0.80	0.81
SEM	3.88%	4.24%	4.07%	4.44%

Standards for the PSA assessments were set by an *a priori* Angoff process. These Angoff standards were sense-checked following psychometric analysis of the February 2022 data, but no post-hoc adjustment to the Angoff standards was deemed necessary. Appendix 3 contains a summary of alternative standard setting considerations using post-hoc analyses.

5. Additional Analyses

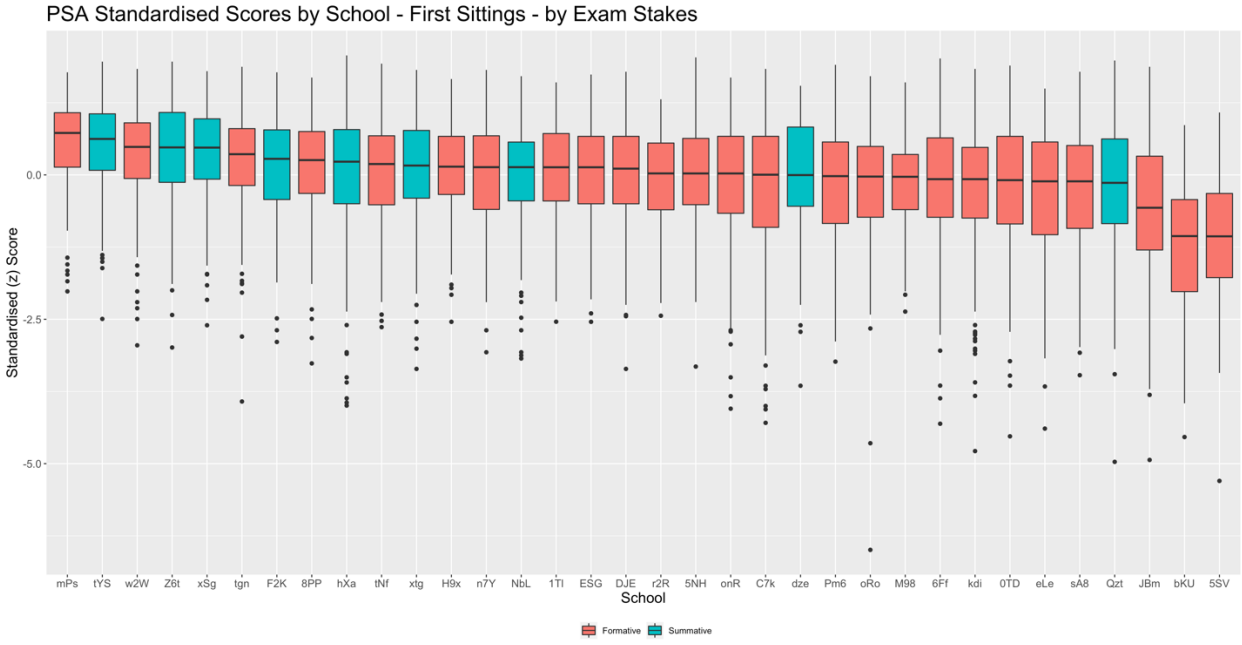
5.1 Summative Use of the PSA

Some medical schools use the PSA as part of their summative assessment portfolio in decision making, in other schools the PSA results do not count towards ranking or awarding. This analysis considers whether there is any association between the stakes of the assessment and student performance (first sitting data).

Scores were higher and failure rates were lower in schools who use the PSA summatively.

	Number in group	Average % Score (95% CI)	t-test for group difference	Failure Rate	χ^2 test for difference in proportion
All Students					
PSA Summative	2605	81.6 (81.3, 81.9)	**t= -10.5 : p< 0.001****	3.22%	**χ^2 = 34.7 ; p< 0.001****
PSA Zero Stakes	5466	79.4 (79.1, 79.6)		6.42%	

The following plot shows standardised scores from first sittings (as in section 3 above) highlighting the schools who use the PSA summatively (blue box plots) or as a zero-stakes assessment (red box plots).

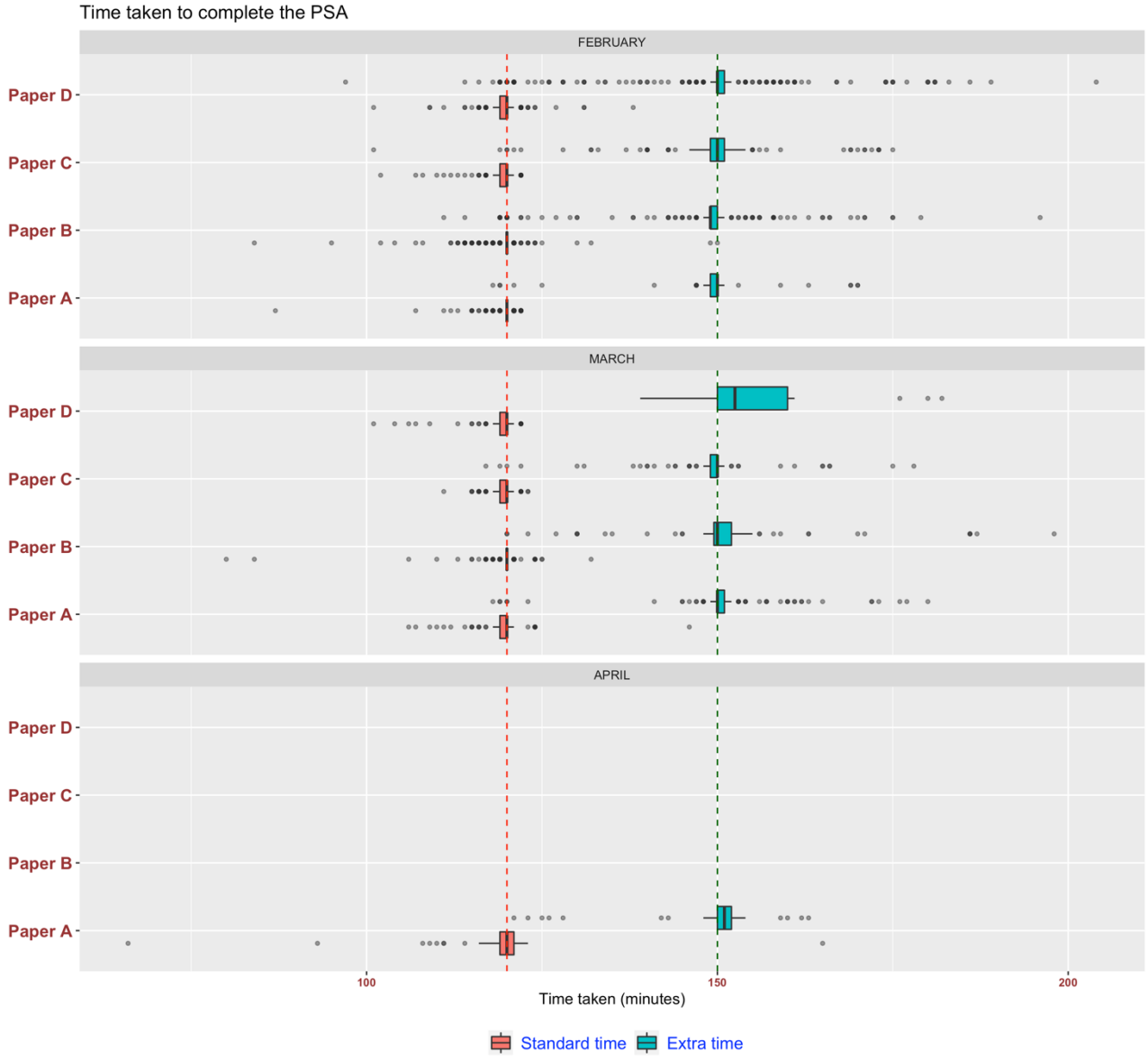


5.2 Examination Timings

Data from the PSA examinations are fully anonymised and only aggregated at the school level. However, data are collected on exam timings, whether students were assigned extra time and whether students sat the PSA in the morning or afternoon. This section considers a basic analysis of whether any of these factors, which we can consider external to the test construct, are related to overall test performance.

Standard-time students ordinarily have 120 minutes and extra-time students ordinarily have 150 minutes to complete the PSA. The plots below show the distributions in time taken to complete the PSA at each test point (all first sitting data). Note that tests at which less than 30 students sat a test paper are excluded from this indicative analysis.

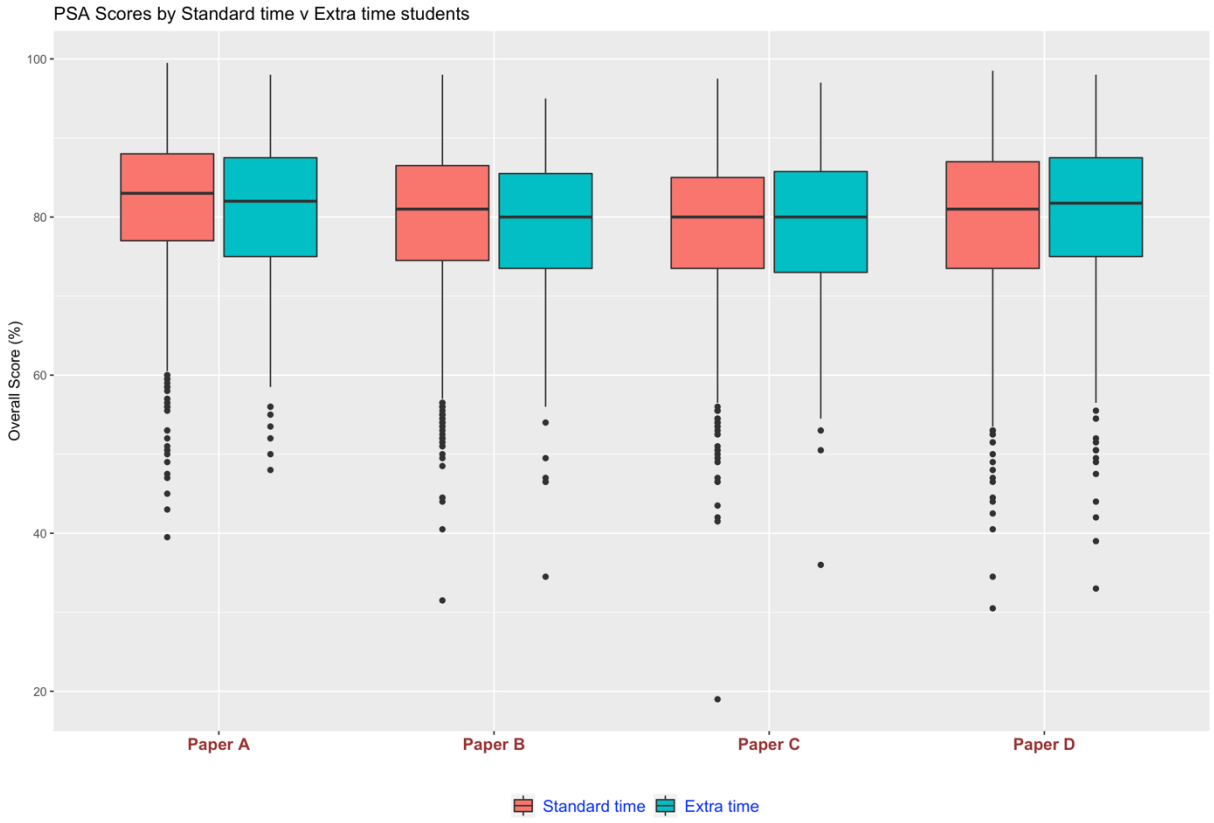
Overall most students took the full amount of time available to complete the assessment.



5.3 Reasonable Adjustments (Extra Time) – First Sitting Data

Overall, students with extra time scored slightly lower than standard time students, though this difference was only nominally significant on paper B. There were no differences in failure rates between standard and extra time students.

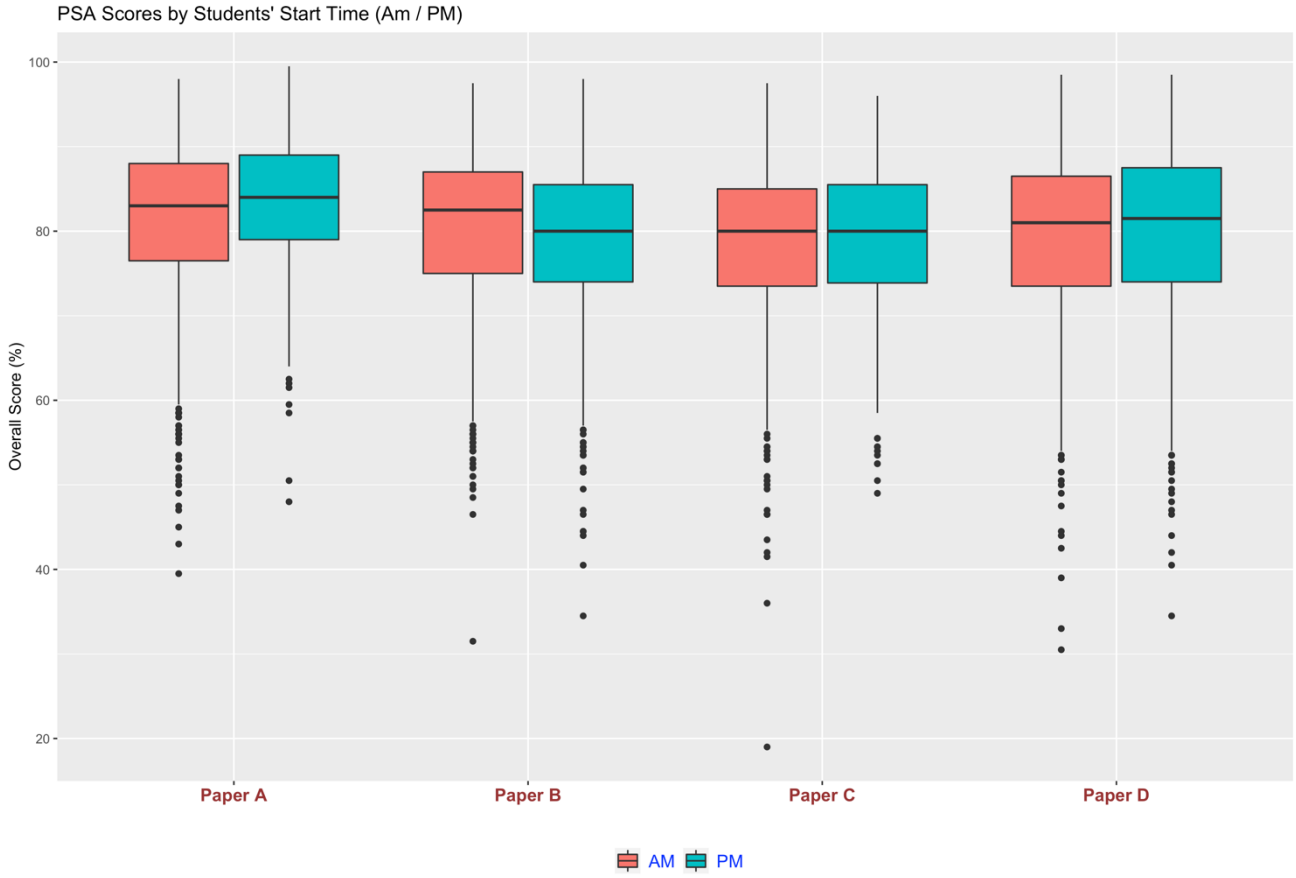
	Number in group	Average % Score (95% CI)	t-test for group difference	Failure Rate	χ^2 test for difference in proportion
All Students					
Standard time	6787	80.2 (80.0, 80.4)	**t= -2.1 : p= 0.036**	6.39%	χ^2 = 2.7 ; p= 0.097
Extra time	1284	79.6 (79.0, 80.1)		5.20%	
Paper A					
Standard time	2008	81.9 (81.5, 82.3)	t= -1.9 : p= 0.055	5.78%	χ^2 = 1.1 ; p= 0.29
Extra time	277	80.8 (79.7, 81.9)		4.18%	
Paper B					
Standard time	1963	79.9 (79.5, 80.3)	**t= -2.2 : p= 0.029**	5.59%	χ^2 = 0.27 ; p= 0.60
Extra time	376	78.7 (77.8, 79.7)		4.79%	
Paper C					
Standard time	1385	78.8 (78.3, 79.3)	t= -0.1 : p= 0.941	5.53%	χ^2 = 0.01 ; p= 0.92
Extra time	235	78.7 (77.5, 80.0)		5.13%	
Paper D					
Standard time	1431	79.5 (79.0, 80.1)	t= 0.7 : p= 0.463	8.08%	χ^2 = 0.19 ; p= 0.662
Extra time	396	80.0 (78.9, 81.1)		7.27%	



5.4 Morning / afternoon sittings

Overall, there was no evidence this year to suggest that start time (morning v afternoon) had any impact on test performance in terms of overall scores or pass/fail rates.

	Number in group	Average % Score (95% CI)	t-test for group difference	Failure Rate	χ^2 test for difference in proportion
All Students					
Morning	4797	80.2 (79.9, 80.5)	t= 1.3 : p= 0.198	5.67%	χ^2 = 1.7 ; p = 0.193
Afternoon	3274	79.9 (79.6, 80.2)		4.98%	
Paper A					
Morning	1792	81.5 (81.1, 81.9)	**t= -3.7 : p< 0.001****	4.91%	**χ^2 = 5.1 ; p < 0.024***
Afternoon	493	83.0 (82.3, 83.7)		2.43%	
Paper B					
Morning	1076	80.4 (79.9, 81.0)	**t= 3.7 : p< 0.001***	5.20%	χ^2 = 0.25 ; p = 0.618
Afternoon	1263	79.0 (78.5, 79.5)		4.67%	
Paper C					
Morning	1076	78.6 (78.0, 79.2)	t= -1.1 : p= 0.268	5.76%	χ^2 = 1.8 ; p=0.176
Afternoon	544	79.1 (78.4, 79.9)		4.04%	
Paper D					
Morning	853	79.3 (78.6, 80.0)	t= -1.5 : p= 0.137	7.74%	χ^2 = 0.13 ; p = 0.720
Afternoon	974	80.0 (79.3, 80.6)		7.19%	



6. Test Equating

In a test administration, where different groups of students are exposed to different test forms, it is important to establish the relative difficulty of the test forms in order that one group is not disadvantaged by sitting a harder paper. For the PSA, standards are set at the item level by a modified Angoff process, so the difficulty of different test papers is mediated by the standard setting.

Nevertheless, a set of anchor items (common to multiple test forms) are embedded in each PSA paper, and post-hoc analysis of the cohort performance on these items, relative to the cohort performance on other (unique) items on the paper, enables the investigation at the aggregated (cohort) level of the relative difficulties of each paper.

In this administration of the PSA, the four papers each contained 36 unique items and 8 pairs of items shared across four papers. The table below shows the mean percentage scores on the full paper, the unique items and each set of anchor items across the four papers. This data is from all first sittings across the year.

	N Items	Paper A	Paper B	Paper C	Paper D
Overall Avg Score	60	81.78	79.67	78.77	79.64
Unique Items Avg	36	84.15	82.36	77.92	81.64
Anchor Pairs:					
AB	8	73.52	74.53		
AC	8	81.18		80.55	
AD	8	79.72			77.99
BC	8		78.88	79.76	
BD	8		72.98		71.98
CD	8			79.96	79.60

The data suggest that there were no significant differences between group performances on anchor sets across the papers – i.e. the spread of students across the four papers were of equal ability on average. Assuming that this finding is generalisable, this suggests that differences in performance on unique items are not influenced by group ability but explain genuine differences in paper difficulty.

These test equating data suggest a rank ordering of difficulty with Paper A being the “easiest” paper, Paper C being the “hardest” and papers B and D being relatively similar in difficulty.

These equating calculations were considered for standard setting purposes following the February tests (the equating data and calculations from February tests and all first sittings is presented in Appendix 5). Importantly, the overall equating calculations from all 2022 first sittings above are in line with the test equating calculations run after the February PSA sitting and used to inform the standard setting at that point in the year. The test equating analyses are confirmatory of the raw score analyses, and did not indicate any requirement for post-hoc adjustment of the Angoff standards for the four papers.

Whilst is should be noted that these test equating calculations are based on average scores and that these calculations are sensitive to the item characteristics in the anchor item sets – including the number of anchor item pairs within tests, the large sample sizes in each test paper makes this equating method fairly robust, and provides a generally good indication of the average group ability differences and average paper difficulty.

7. Item Analysis

The following analyses present item-level data from each paper, to highlight any items which are worthy of review based on the performance data. A classical item analysis was run on the full test papers, and a single parameter RASCH model was fitted to the dichotomous single best answer (SBA) items for comparison purposes. The full item analyses tables for each paper are in Appendix 6. This section highlights questions which might merit a closer look.

Facility (Difficulty)

Items in a test which score 100% or items which were not answered correctly by anyone, do not add anything to the test in terms of discriminating between good and poor students, and should generally be avoided. Very low facility items should be investigated for answer keying errors or content irrelevance. Items where the cohort average score is lower than the Angoff standard score for the item also merit a closer look as the standard may have been set too high for such questions.

Paper A
17 items had facility over 90% : 16 uniques, one AC anchor
1 unique item (MAN.UK.01.4871) had facility under 30%
4 items had a mean score below the Angoff score
Paper B
9 items had facility over 90% : 8 uniques, one AB anchor
1 unique item (MAN.UK.01.4911) had facility under 30%
4 items had a mean score below the Angoff score
Paper C
15 items had facility over 90% : 13 uniques, 1 AC anchor, 1 BC anchor
0 items had facility under 30%
4 items had a mean score below the Angoff score
Paper D
10 unique items had facility over 90%
0 items had facility under 30%
1 item had a mean score below the Angoff score

Discrimination

Discrimination is the degree to which success on an individual item corresponds to success on the whole test.

The discrimination index is an index of an item's effectiveness at discriminating those who know the content from those who don't. It is computed from equal-sized high and low scoring groups on the test (for this analysis the top and bottom 1/3 of the cohort). The number of successes by the low group on the item is subtracted from the number of successes by the high group, and this difference is divided by the size of a group. The range of this index is +1 to -1. A discrimination index close to zero suggests that the item does not discriminate well between high and low scoring students (this may be seen in very easy items where nearly all students are successful). A negative value suggests that the lower scoring students do better on the item than the higher scoring students, and reasons for this should be explored.

The corrected item-total correlation (CITC) is the correlation between responses to a particular item and scores on the total test, without that item. The ranges for this metric are +1 to -1. A negative CITC suggests that the item is inversely correlated with the rest of the items in the test – i.e. candidates who perform well on the item perform poorly on the test overall. Reasons for this should usually be explored.

Paper A
No items showed negative CITC or discrimination.
Paper B
1 unique item (MAN.UK.01.4911) had negative CITC and low discrimination.
Paper C
No items showed negative CITC or discrimination.
Paper D
No items showed negative CITC or discrimination.

Factor Loading

The relationship of each variable to the underlying factor is expressed by the so-called factor loading and can be interpreted like a correlation coefficient with the factor. Items in a test with a factor loading of less than 0.1 do not correlate well with the underlying latent trait.

Paper A
No items had a factor loading of less than 0.1.
Paper B
1 unique item (MAN.UK.01.4911) had a negative factor loading.
Paper C
1 unique item had a factor loading of less than 0.1.
Paper D
1 unique item had a factor loading of less than 0.1.

Internal Consistency (Cronbach's α if item deleted)

An item's contribution to internal consistency is measured by estimating alpha with that item removed from the set. The result is a statistic called alpha-if-item-deleted (AID). Where AID is higher than the overall alpha, the assessment's reliability may be improved by removing the item.

No items impacted adversely on the test reliability to any significant degree in any test papers.

8. RASCH Analysis

The RASCH model uses all information from persons and items and fits items in a test on a single latent dimension. The model requires the data to have certain properties in order to work well, one of which is an assumption of unidimensionality already mentioned. Another related assumption is that success or failure on any individual item should not depend on success or failure on any item (conditional independence). Often in longer tests which cover multiple curriculum areas these assumptions are violated.

In the following analyses a simple one parameter RASCH model is used which is only suitable for dichotomous item data so is only applied to the SBA items of the PSA (not the PWS and REV items).

Unlike the CTT discrimination statistics, the RASCH model fit statistics assume that all items in a test discriminate in the same way, and are based on model residuals rather than on comparison of groups of student performance. The RASCH model produces various item (and person) statistics which can be considered to investigate the performance and validity of items within tests.

Outfit and Infit statistics from the RASCH model

Outfit is an "unweighted fit" statistic. The outfit can be thought of as an "outlier sensitive" measure and is sensitive to extreme departures from model expectations. For example, an extreme departure from model expectations would occur when an otherwise high-achieving student provided an incorrect response to a very easy item, or when an otherwise low-achieving student provided a correct response to a very difficult item.

Infit is an "information-weighted fit" statistic. The infit statistic can be thought of as an "inlier sensitive" measure and is sensitive to less-extreme unexpected responses compared to outfit. Examples of less-extreme unexpected responses include a student providing an incorrect response to an item that is just below their achievement level, or a student providing a correct response to an item that is just above their achievement level.

Mean-square (MSQ) fit statistics are reported which show the amount of variation in how the data fit the model. Commonly agreed-upon principles for interpreting these statistics.

- Expected value is 1 when data fit the model
- Less than 1: Responses are too predictable (possibly redundant – data overfit the model)
- Greater than 1: Responses are too noisy (there is too much variation to suggest that the estimate is a good representation of the response pattern)
- Some variation is expected, but noisy responses (larger than 1) are usually considered more cause for concern than muted responses (less than 1)
- Generally acceptable values for MSQ Infit and Outfit lie in the range 0.7 – 1.3 (Wright & Linacre, 1994)

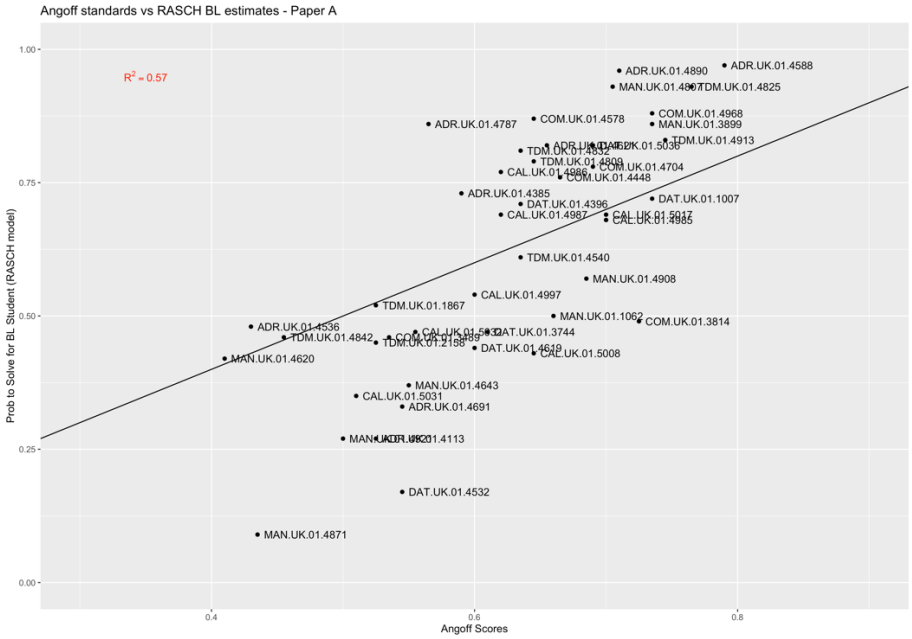
Two items on paper A, and one item each on paper B, paper C and paper D had slightly low Outfit, but all were very easy items with facility near 100% so likely moderately overfit the model.

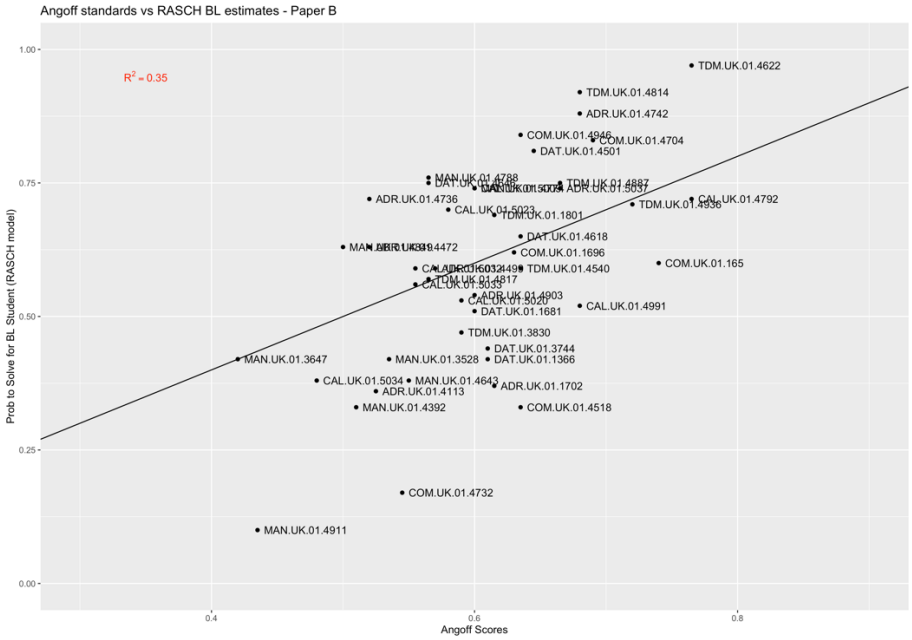
Appendix 7 provides further analysis of the RASCH fit statistics for reference. The item characteristic curves (ICCs) for each item in each paper, showing the probability of a correct response as a function of the ability of persons is also reported in Appendix 7. The final section of this appendix shows a comparison between item facility from the classical testing framework against beta difficulty estimates from the RASCH model.

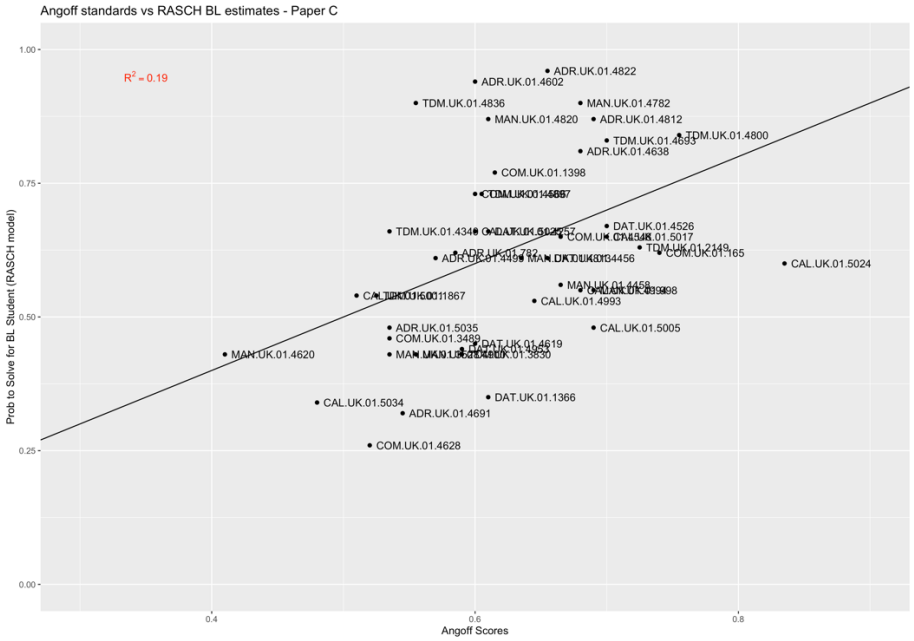
RASCH modelling of Borderline Students

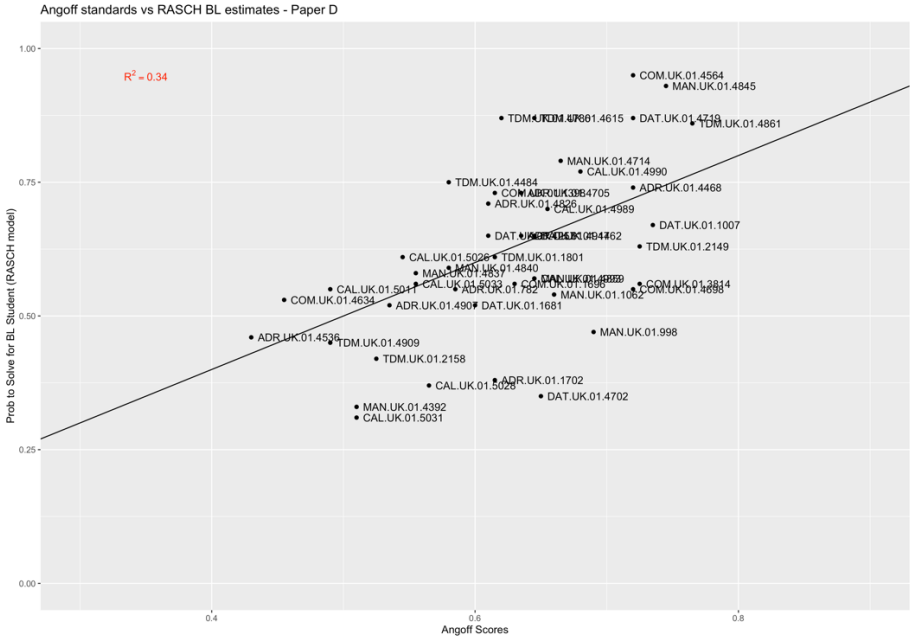
Because the RASCH model separates the estimation of item difficulty out from person ability and places them on the same latent scale, a nice property of the model is that it enables estimation of expected probability to solve items for each person ability level. The following plots use the RASCH item and person predictions to estimate the proportion of borderline students who would answer the item correctly, plotted against the Angoff score for that item. Ideally items should be on or near the straight line. Items below the line may suggest that the Angoff score is too high, items above, that the Angoff score is too low.

It should be noted that for the probabilistic RASCH model there is no provision in the model for random guessing, suggesting that a person of very low ability will always get an item wrong. In actuality for 5-item SBAs as in the PSA, low-ability individuals have a random probability of success of 0.2 on any item. In practice for this assessment this should not be a big issue, as most guessing, even amongst borderline candidates will be informed guessing. However, this aspect of the RASCH model may deflate the probability of success especially at the lower end of the scale.






Angoff standards vs RASCH BL estimates - Paper C
R² = 0.19
ADR.UK.01.4822
ADR.UK.01.4602
TDM.UK.01.4836
MAN.UK.01.4782
MAN.UK.01.4820
ADR.UK.01.4812
TDM.UK.01.4800
TDM.UK.01.4693
ADR.UK.01.4638
COM.UK.01.1398
DAT.UK.01.4526
TDM.UK.01.2149
COM.UK.01.165
CAL.UK.01.5024
MAN.UK.01.4468
CAL.UK.01.4993
ADR.UK.01.5035
CAL.UK.01.5005
COM.UK.01.3489
DAT.UK.01.4619
MAN.UK.01.4620
CAL.UK.01.5034
DAT.UK.01.1366
ADR.UK.01.4691
COM.UK.01.4628
Prob to Solve for BL Student (RASCH model)
Angoff Scores
1.00
0.75
0.50
0.25
0.00
0.4
0.6
0.8


Angoff standards vs RASCH BL estimates - Paper D
R² = 0.34
COM.UK.01.4564
MAN.UK.01.4845
TDM.UK.01.4615
DAT.UK.01.4719
TDM.UK.01.4861
MAN.UK.01.4714
CAL.UK.01.4990
TDM.UK.01.4484
ADR.UK.01.4468
ADR.UK.01.4826
CAL.UK.01.4989
DAT.UK.01.1007
TDM.UK.01.2149
CAL.UK.01.5026
TDM.UK.01.1801
MAN.UK.01.4840
COM.UK.01.3814
CAL.UK.01.5011
COM.UK.01.4634
MAN.UK.01.1062
ADR.UK.01.4907
DAT.UK.01.1681
MAN.UK.01.998
ADR.UK.01.4536
TDM.UK.01.4909
TDM.UK.01.2158
ADR.UK.01.1702
CAL.UK.01.5028
DAT.UK.01.4702
MAN.UK.01.4392
CAL.UK.01.5031
Prob to Solve for BL Student (RASCH model)
Angoff Scores
1.00
0.75
0.50
0.25
0.00
0.4
0.6
0.8

Appendix 1 – Paper Allocation by School by Sitting

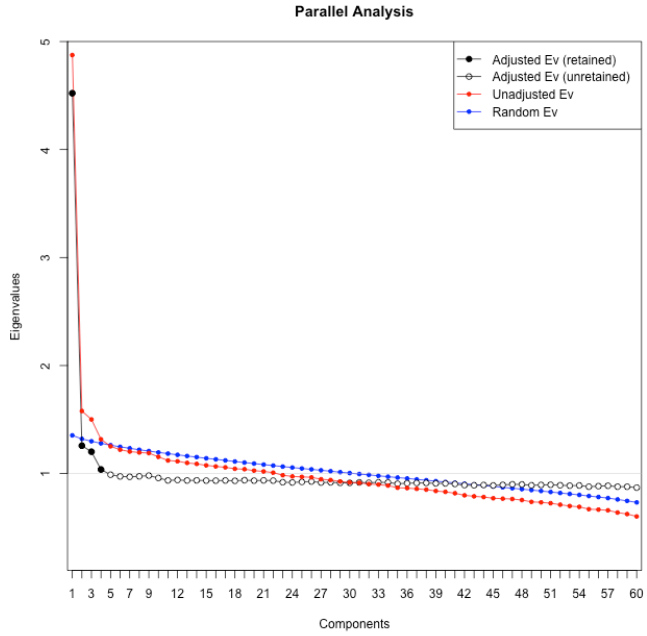
See separate appendix.

Appendix 2 – Factor Analysis for Dimensionality

The key concept of factor analysis in test analyses is that multiple test items have similar patterns of responses because they are all associated with a latent (not directly measured) trait, such as "medical knowledge". Every factor analysis begins by assuming the same number of factors as there are items in a test. Each factor captures a certain amount of the overall variance in the observed variables. The reported eigenvalues are a measure of how much of the variance of the observed variables a factor explains. Any factor with an eigenvalue >1 explains more variance than a single observed variable and is generally retained as a factor (or dimension) of the test. The larger the eigenvalue, the greater the importance of the factor.

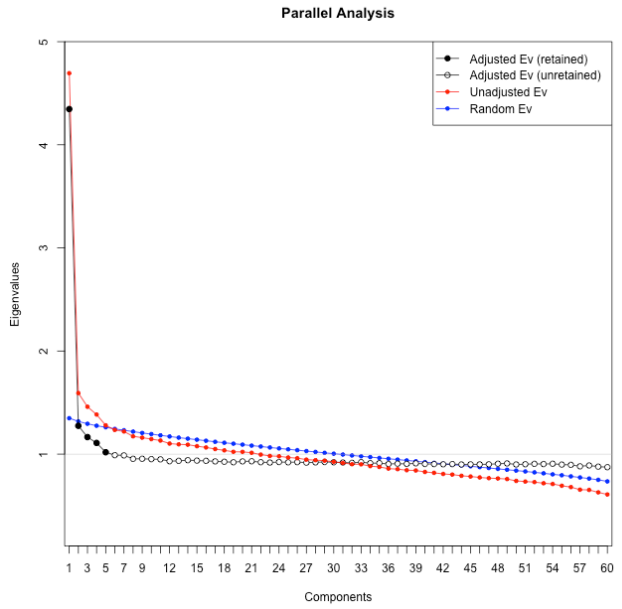
For each PSA paper, a parallel analysis was run to estimate the number of dimensions or factors present in the test. The following scree plots and output from R show the parallel analysis results.

Paper A: Parallel Analysis (Principle Components Analysis)



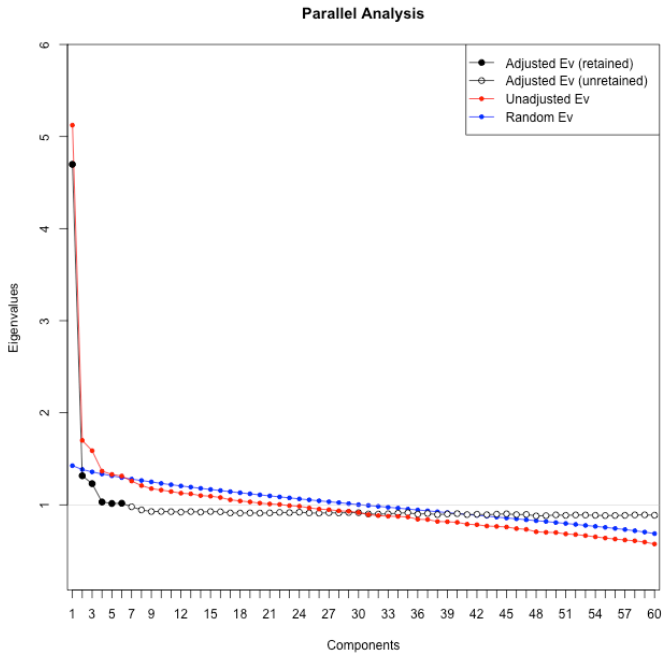

```
Results of Horn's Parallel Analysis for
component retention
1000 iterations, using the 95 centile estimate

-------------------------------------------------
Component   Adjusted     Unadjusted    Estimated
            Eigenvalue   Eigenvalue    Bias
-------------------------------------------------
1           4.521125     4.874817      0.353691
2           1.257969     1.578996      0.321027
3           1.202240     1.501153      0.298913
4           1.037429     1.317204      0.279775
-------------------------------------------------

Adjusted eigenvalues > 1 indicate dimensions to
retain.
(4 components retained)
```

Paper B: Parallel Analysis (Principle Components Analysis)



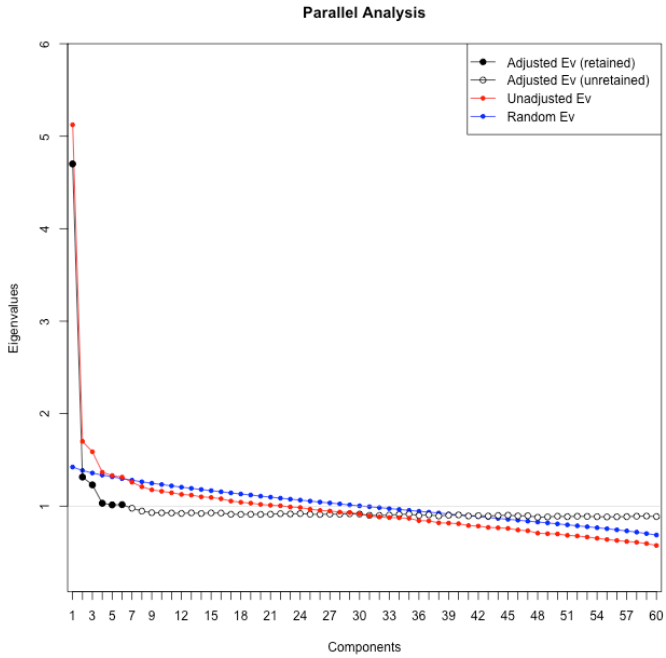

```
Results of Horn's Parallel Analysis for
component retention
1000 iterations, using the 95 centile estimate

-------------------------------------------------
Component   Adjusted     Unadjusted    Estimated
            Eigenvalue   Eigenvalue    Bias
-------------------------------------------------
1           4.346052     4.694571      0.348519
2           1.274683     1.592084      0.317400
3           1.165066     1.459364      0.294297
4           1.108444     1.383821      0.275377
5           1.018488     1.278959      0.260471
-------------------------------------------------

Adjusted eigenvalues > 1 indicate dimensions to
retain.
(5 components retained)
```

Paper C: Parallel Analysis (Principle Components Analysis)




```
Results of Horn's Parallel Analysis for
component retention
1000 iterations, using the 95 centile estimate

-----------------------------------------------
Component   Adjusted     Unadjusted    Estimated
            Eigenvalue   Eigenvalue    Bias
-----------------------------------------------
1           4.697768     5.123362      0.425594
2           1.315928     1.699994      0.384066
3           1.230205     1.587960      0.357754
4           1.030848     1.366365      0.335516
5           1.015032     1.330388      0.315356
6           1.016861     1.313270      0.296409
-----------------------------------------------

Adjusted eigenvalues > 1 indicate dimensions to
retain.
(6 components retained)
```

Paper D: Parallel Analysis (Principle Components Analysis)



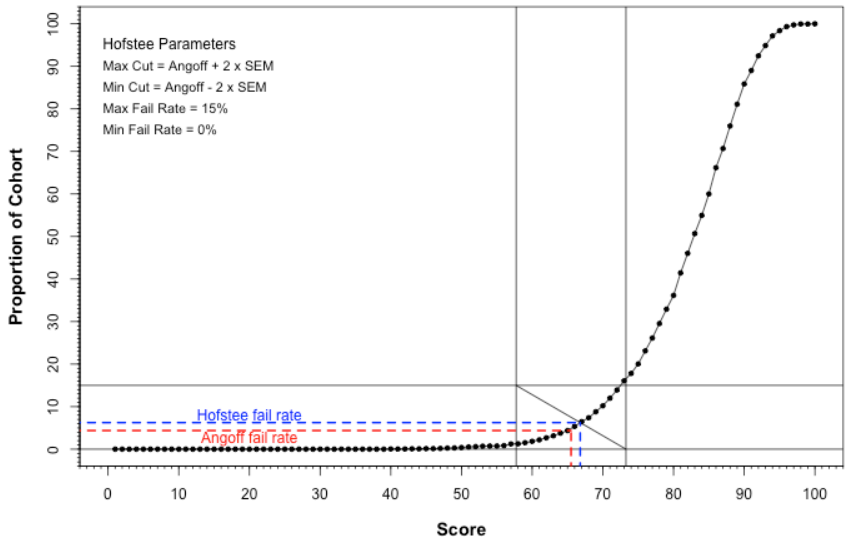

```
Results of Horn's Parallel Analysis for
component retention
1000 iterations, using the 95 centile estimate

-----------------------------------------------
Component   Adjusted     Unadjusted    Estimated
            Eigenvalue   Eigenvalue    Bias
-----------------------------------------------
1           4.700673     5.123362      0.422689
2           1.314565     1.699994      0.385429
3           1.230010     1.587960      0.357950
4           1.030960     1.366365      0.335405
5           1.013655     1.330388      0.316733
6           1.015761     1.313270      0.297509
-----------------------------------------------

Adjusted eigenvalues > 1 indicate dimensions to
retain.
(6 components retained)
```

Appendix 3 – Standard Setting – First Sitting Data

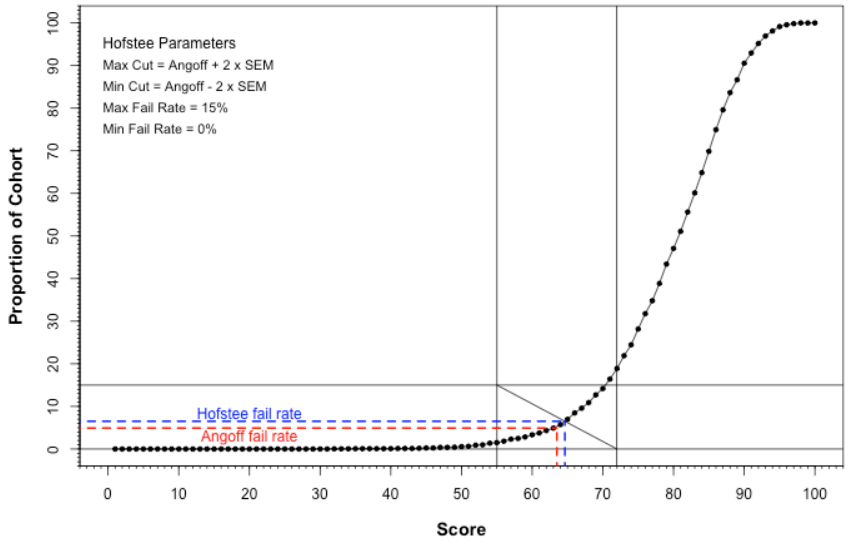
The Angoff standards were applied for all four papers with post-hoc adjustment. The table below considers theoretical alternative standards and failure rates using adjustments for SEMs and a post-hoc Hofstee for each paper. These analyses use first sitting data from all test dates.

	Paper A		Paper B		Paper C		Paper D	
	Cut-Score	Fail Rate (n)	Cut-Score	Fail Rate (n)	Cut-Score	Fail Rate (n)	Cut-Score	Fail Rate (n)
Angoff Pass Mark (PM)	**65.5%**	**4.4% (100)**	**63.5%**	**4.9% (115)**	**63%**	**5.2% (84)**	**64%**	**7.4% (136)**
PM + 1 SEM	69.38%	8.8% (201)	67.74%	10.3% (242)	67.07%	11.2% (181)	68.44%	13.2% (242)
PM + 2 SEM	73.27%	16.1% (367)	71.98%	17.5% (410)	71.14%	18.8% (305)	72.87%	21.9% (400)
PM - 1 SEM	61.62%	2.4% (54)	59.26%	2.8% (66)	58.93%	2.8% (46)	59.56%	4.8% (87)
PM - 2 SEM	57.73%	1.2% (28)	55.02%	1.8% (42)	54.86%	1.7% (28)	55.13%	2.6% (47)
Hofstee Pass Mark	66.8%	5.7% (131)	64.63%	6.3% (148)	63.98%	6.3% (102)	63.72%	7.4% (136)

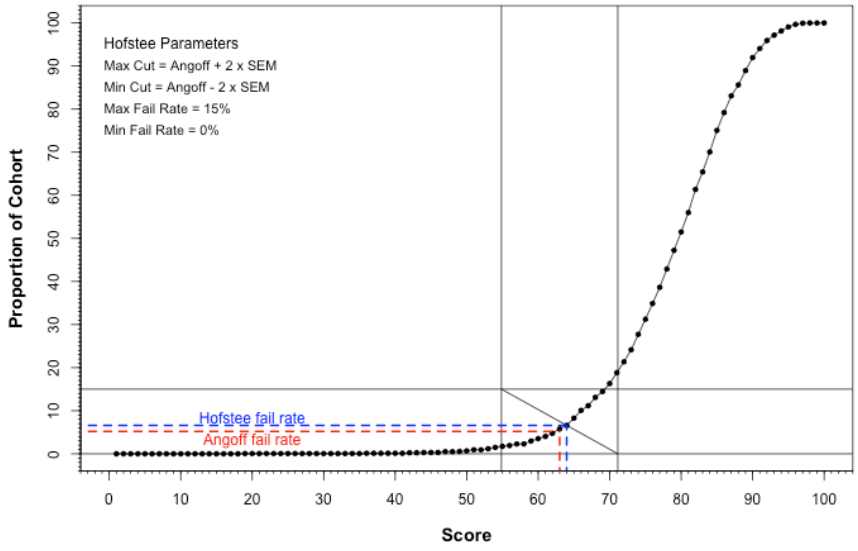
Paper A: Frequency Distribution with Hofstee



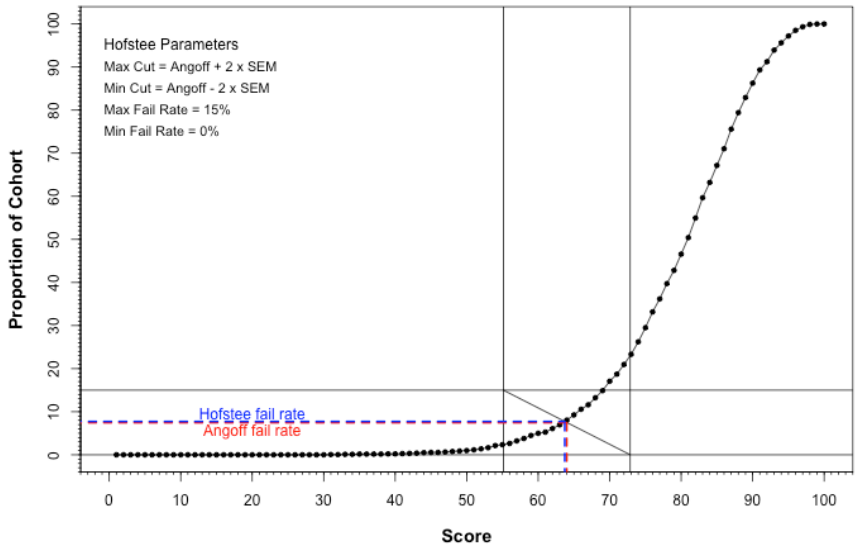
Paper B: Frequency Distribution with Hofstee



Paper C: Frequency Distribution with Hofstee



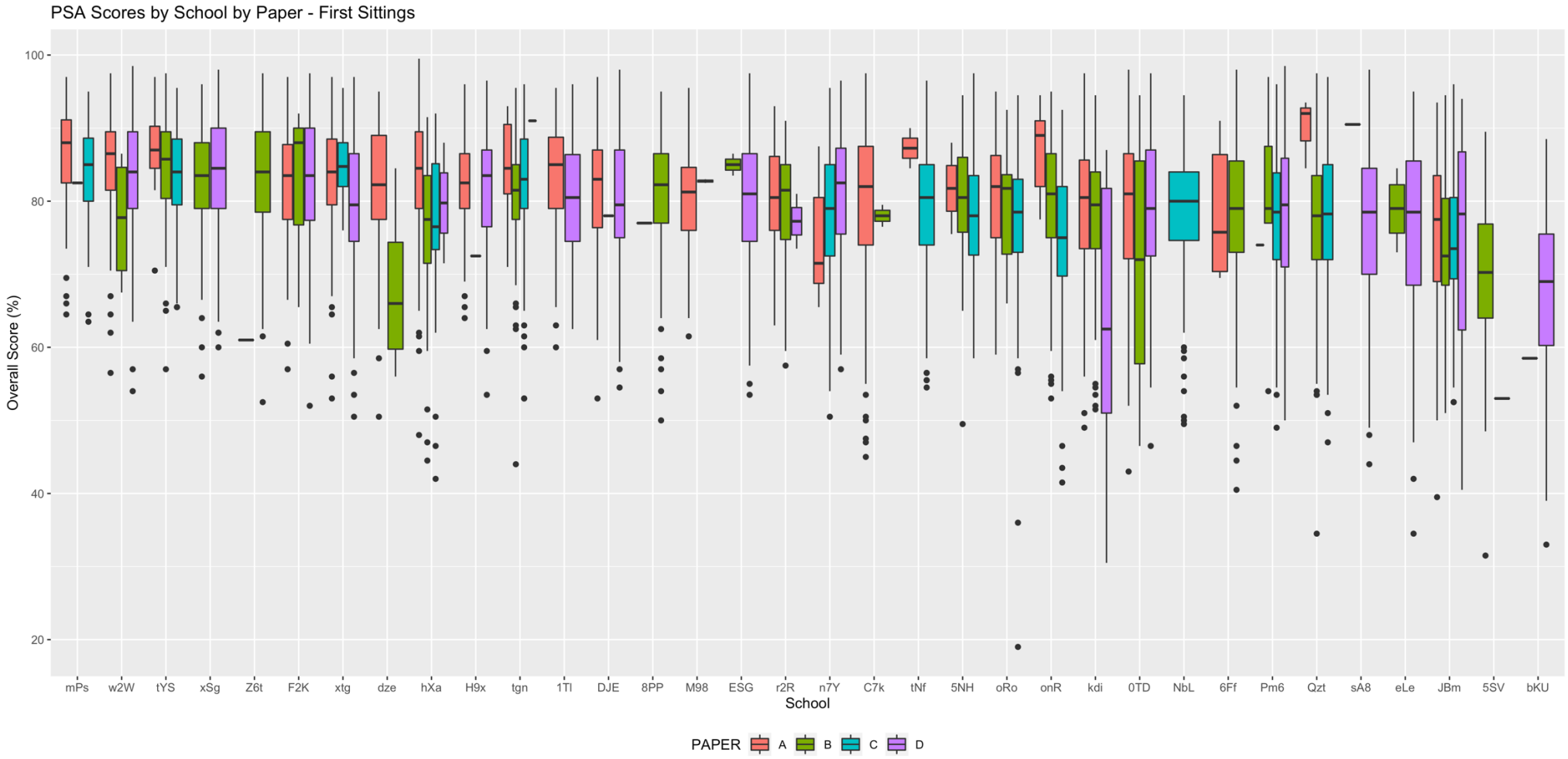
Paper D: Frequency Distribution with Hofstee

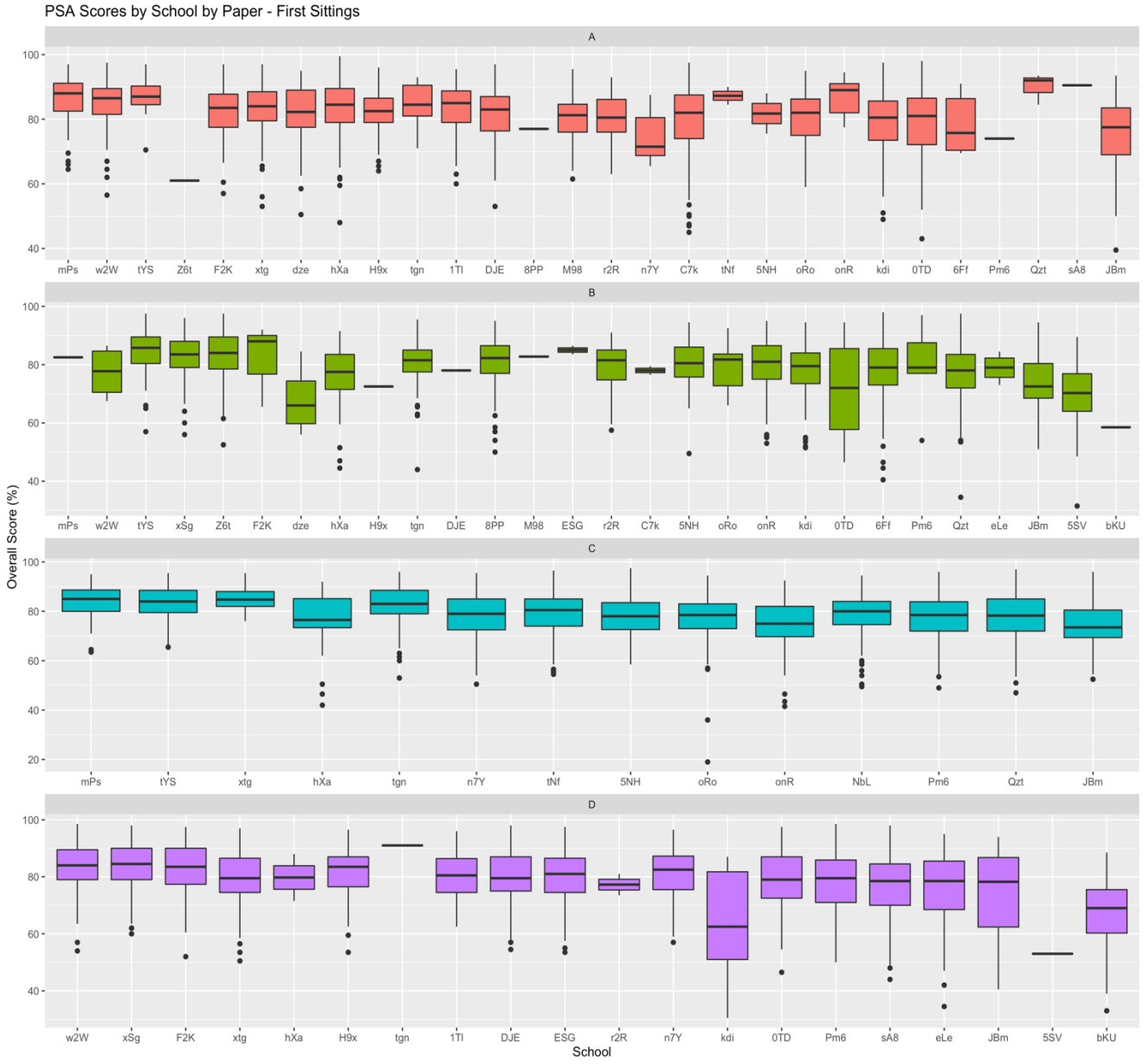


Appendix 4 – Total test scores by school split by paper

The following plots show the first sitting score distributions on each paper by school (all dates). Note that occasionally schools administer a sitting / paper to only one or a few students (see Appendix 1 for numbers), and this clearly has implications in these analyses.

The first plot shows distributions of scores by paper grouped by school, and the second set of plots show the disaggregated data by test paper.




PSA Scores by School by Paper - First Sittings
A
B
C
D
Overall Score (%)
School
mPs w2W tYS Z6t F2K xtg dze hXa H9x tgn 1TI DJE 8PP M98 r2R n7Y C7k tNf 5NH oRo onR kdi 0TD 6Ff Pm6 Qzt sA8 JBm
mPs w2W tYS xSg Z6t F2K dze hXa H9x tgn DJE 8PP M98 ESG r2R C7k 5NH oRo onR kdi 0TD 6Ff Pm6 Qzt eLe JBm 5SV bKU
mPs tYS xtg hXa tgn n7Y tNf 5NH oRo onR NbL Pm6 Qzt JBm
w2W xSg F2K xtg hXa H9x tgn 1TI DJE ESG r2R n7Y kdi 0TD Pm6 sA8 eLe JBm 5SV bKU

Appendix 5: Test equating calculations

The relative difficulty of each paper can be estimated by using the average performance on the anchor items to estimate the average test group ability differences and then using these group ability estimates along with the average scores on the unique items in each paper to estimate expected scores for each group on each paper. The two sets of calculations below show i) the overall equating from all 2022 first sitting data, and ii) the equating calculations from the February 2022 PSA sittings which were used in the standard setting review following the February sittings. In the tables below, non-shaded cells are average scores from the data (not estimated). Yellow cells are calculated estimates for overall group scores on a paper, given ability estimates from performance on anchor items. Blue shaded cells are group ability estimates between papers.

i) Test equating calculations All 2022 PSA First Sittings

Unique Items (no embedded anchors)

	A	B	C	D
Group 1	84.15			
Group 2		82.36		
Group 3			77.92	
Group 4				81.64

Average Score

	A	B	C	D
Group 1	81.78	78.59	79.39	81.41
Group 2	82.90	79.67	77.90	80.75
Group 3	81.15	80.56	78.77	80.00
Group 4	80.01	78.58	78.42	79.64
Average ->	81.46	79.35	78.62	80.45

Anchor Items (all items)

	AB	AC	AD	BC	BD	CD
Group 1	73.52	81.18	79.72			
Group 2	74.53			78.88	72.98	
Group 3		80.55		79.76		79.96
Group 4			77.99		71.98	79.6

Anchor Group Factors

	Paper A	Paper B	Paper C	Paper D
Group 1		0.986	1.008	1.022
Group 2	1.014		0.989	1.014
Group 3	0.992	1.011		1.005
Group 4	0.978	0.986	1.00	

equating factors % difference between groups on papers

	B -> A	C-> A	D -> A	C -> B	D -> B	D -> C
Group 1	3.2	2.4	0.4	-0.8	-2.8	-2.0
Group 2	3.2	5.0	2.2	1.8	-1.1	-2.8
Group 3	0.6	2.4	1.1	1.8	0.6	-1.2
Group 4	1.4	1.6	0.4	0.2	-1.1	-1.2
Average % Difference (equating factor)	2.3	3.3	1.2	0.9	-1.1	-2.0

ii) Test equating calculations from February 2022 PSA sittings (used in standard setting review following the February sittings)

Unique Items (no embedded anchors)

	A	B	C	D
Group 1	84.29			
Group 2		82.42		
Group 3			77.61	
Group 4				81.86

Average Score

	A	B	C	D
Group 1	82.18	78.29	80.03	82.26
Group 2	83.80	79.83	77.66	80.72
Group 3	80.57	80.65	78.46	79.27
Group 4	79.99	79.19	79.26	80.07
Average ->	81.64	79.49	78.85	80.58

Anchor Items (all items)

	AB	AC	AD	BC	BD	CD
Group 1	74.08	81.77	80.81			
Group 2	75.54			78.6	73.21	
Group 3		80.17		79.41		79.76
Group 4			78.66		72.62	80.57

Anchor Group Factors

	Paper A	Paper B	Paper C	Paper D
Group 1		0.981	1.020	1.027
Group 2	1.020		0.990	1.008
Group 3	0.980	1.010		0.990
Group 4	0.973	0.992	1.010	

equating factors % difference between groups on papers

	B -> A	C-> A	D -> A	C -> B	D -> B	D -> C
Group 1	3.9	2.2	-0.1	-1.7	-4.0	-2.2
Group 2	4.0	6.1	3.1	2.2	-0.9	-3.1
Group 3	-0.1	2.1	1.3	2.2	1.4	-0.8
Group 4	0.8	0.7	-0.1	-0.1	-0.9	-0.8
Average % Difference (equating factor)	2.6	3.5	1.4	0.9	-1.2	-2.0

Summary table of equating calculations input data from February 2022 Sittings

	N Items	Paper A	Paper B	Paper C	Paper D
Overall Avg Score	60	82.18	79.83	78.46	80.07
Unique Items Avg	36	84.29	82.42	77.61	81.86
Anchor Pairs:					
AB	8	74.08	75.54		
AC	8	81.77		80.17	
AD	8	80.81			78.66
BC	8		78.6	79.41	
BD	8		73.21		72.62
CD	8			79.76	80.57

Appendix 6 - Item Analysis

See separate appendix.

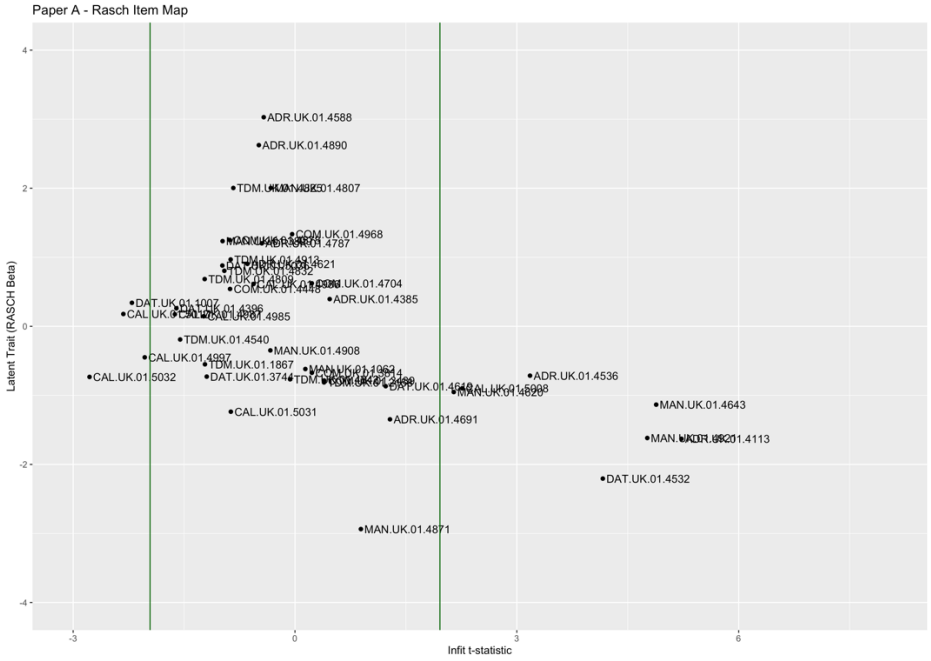
Appendix 7 – RASCH Analysis

RASCH Model Item-Maps

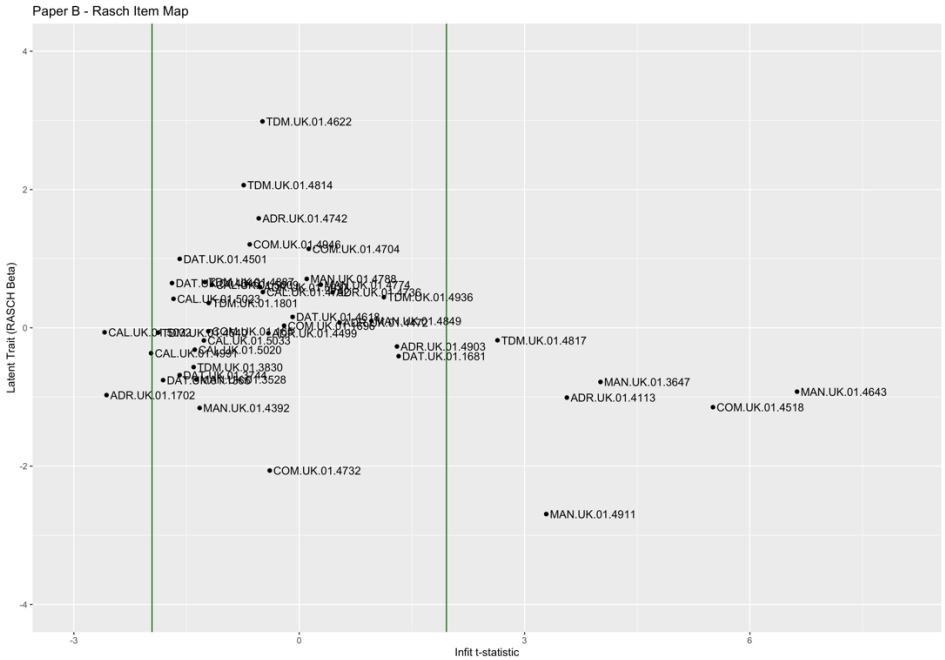
Standardised fit statistics for infit and outfit are t-tests of the hypothesis, "do the data fit the model perfectly?". They are reported as standardised z-scores (means of 0, sd of 1). They show the improbability of the data, i.e., its significance, if the data actually did fit the model. Similar to the MSQ fit statistics, values of less than 0 indicates too predictable (overfit); greater than 0 indicates lack of predictability (underfit). Standardised t-statistics should ideally lie within +/- 1.96 (sds) of the expected score.

The following item-map plots show the results of these t-tests for each paper, showing items outside of the standardised thresholds for indicative purposes only and for comparison with the CTT item analyses (bearing in mind that the MSQ outfit and infit statistics discussed in the body of the report are Ok for all these items).

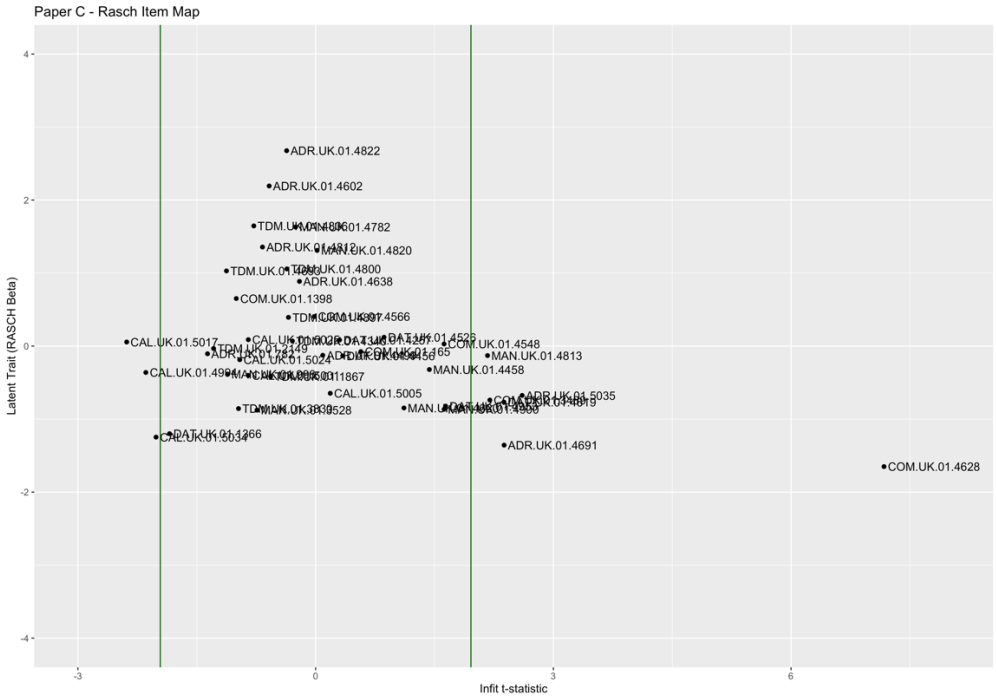
Paper A Item Map from RASCH model



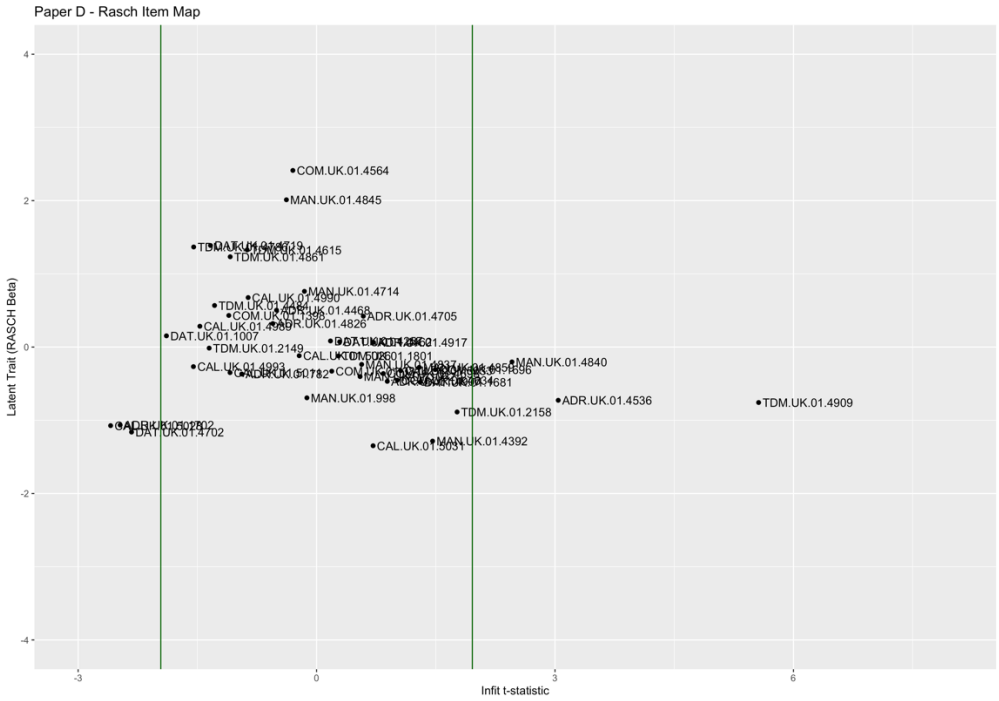
Paper B Item Map from RASCH model



Paper C Item Map from RASCH model



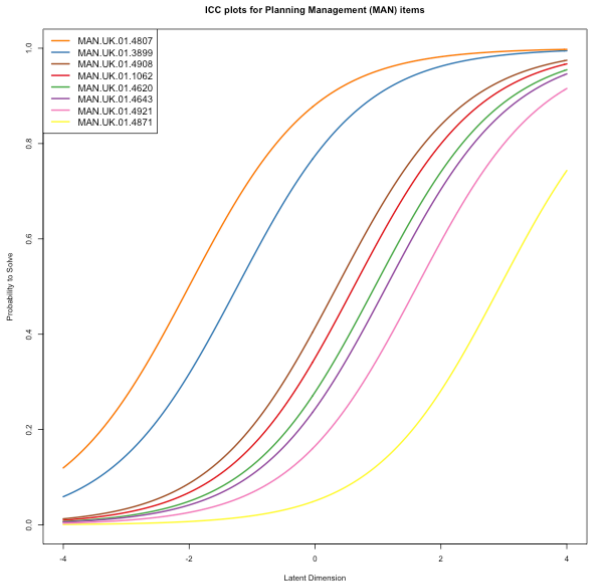
Paper D Item Map from RASCH model

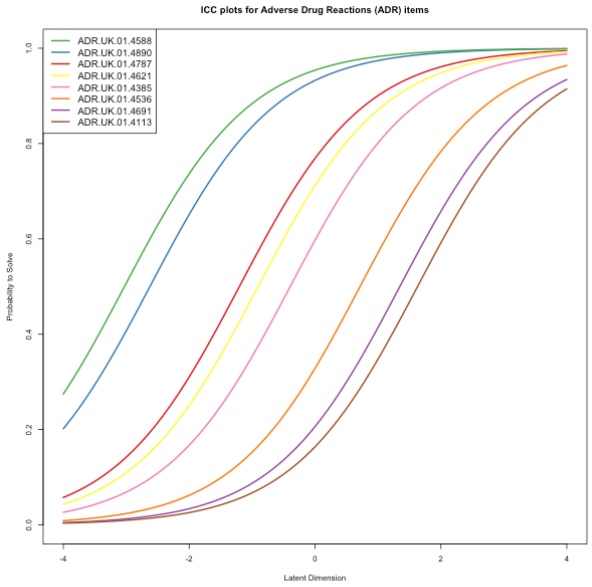


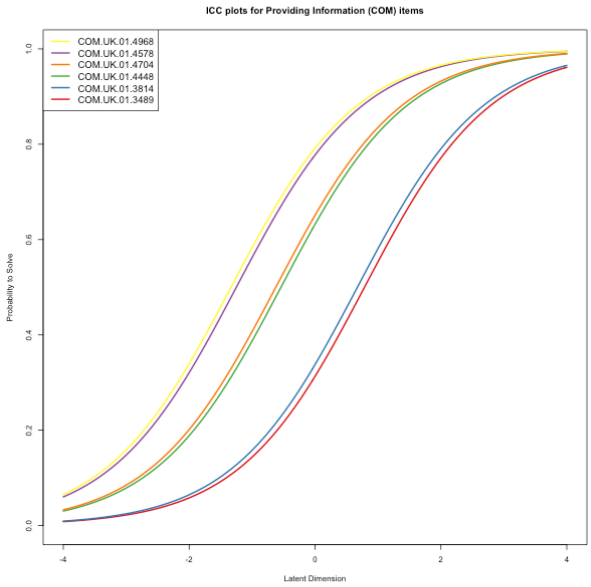
Item Characteristic Curves (Item Response Functions)

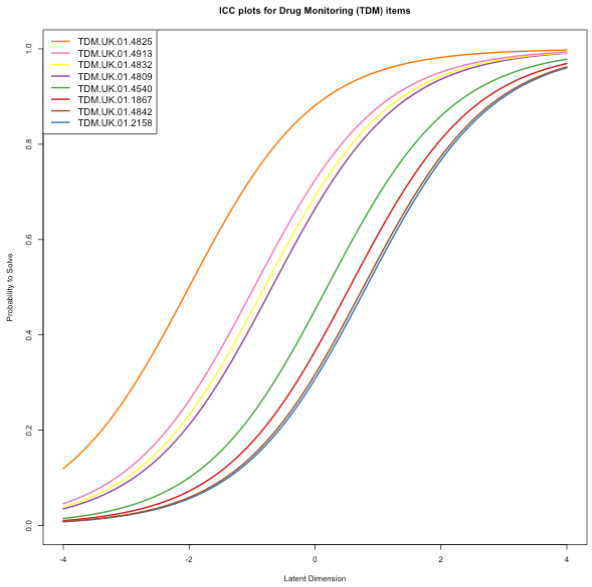
The following ICCs show the probability of a correct response as a function of the ability of the test takers. The plots are grouped by the PSA paper sections. The leftmost (higher) ICCs are the easier items, the rightmost items in the same figures are the most difficult items.

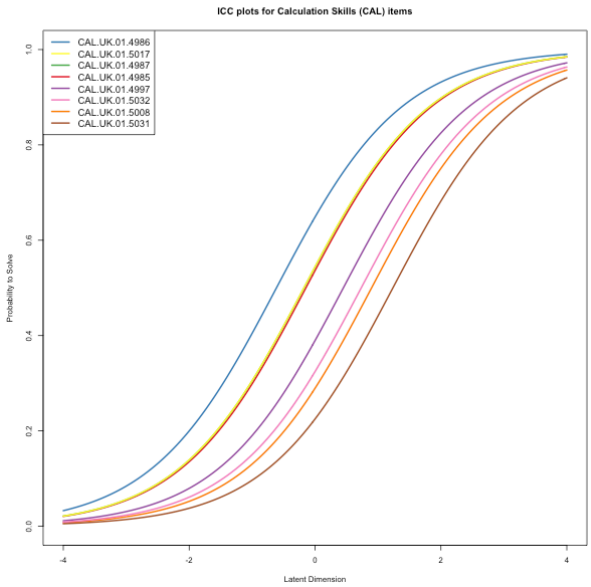
Paper A – RASCH model ICCs

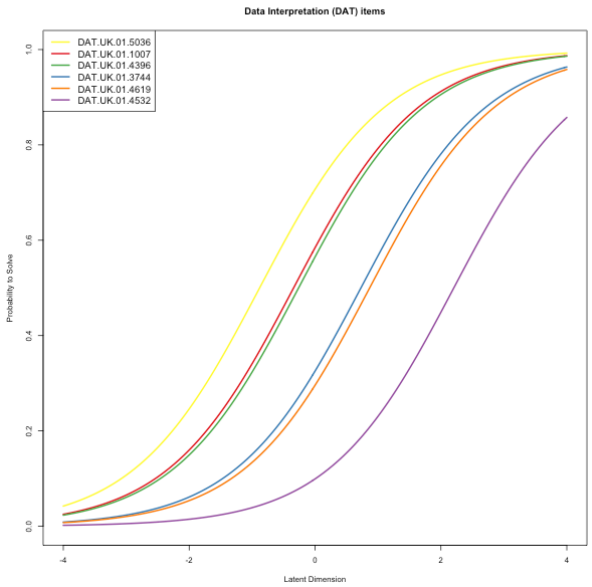




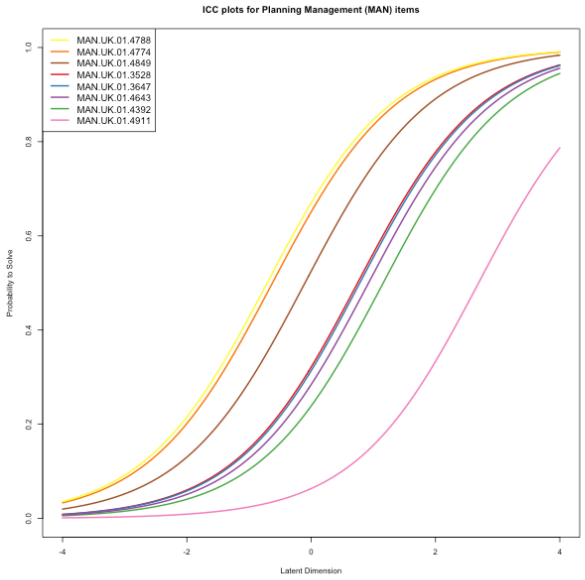


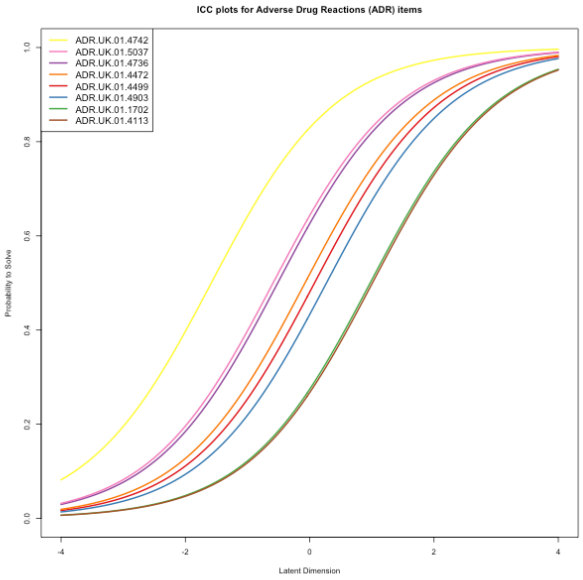


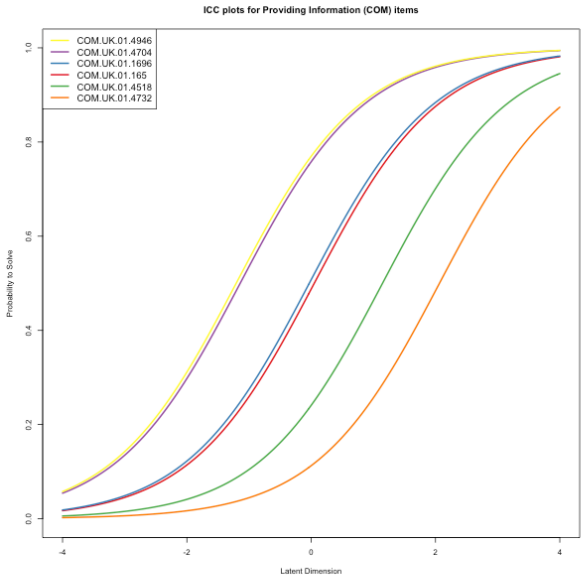


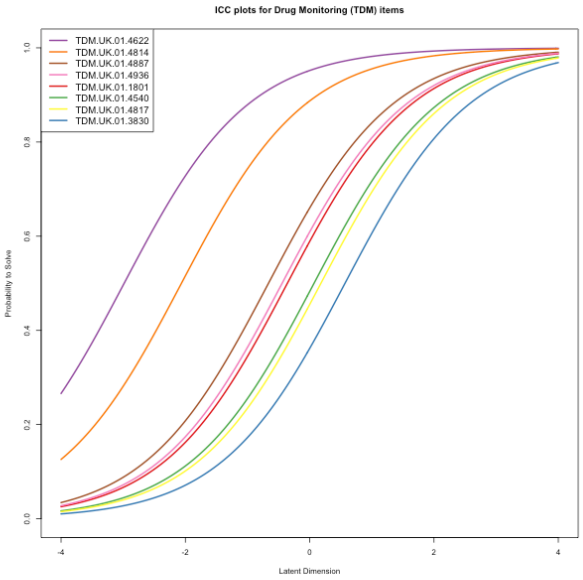


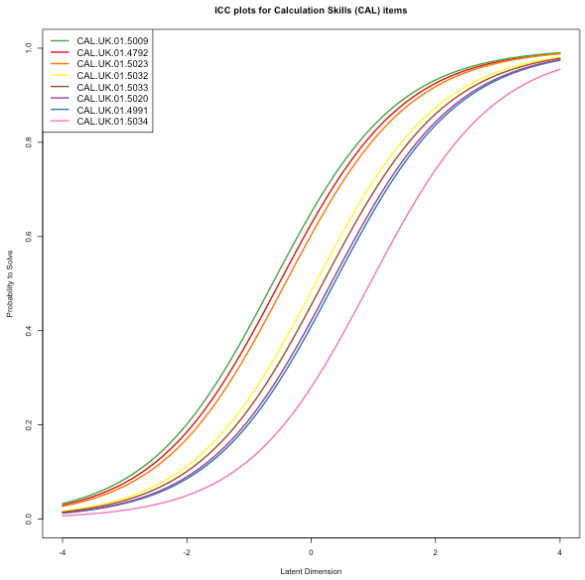
Paper B – RASCH model ICCs

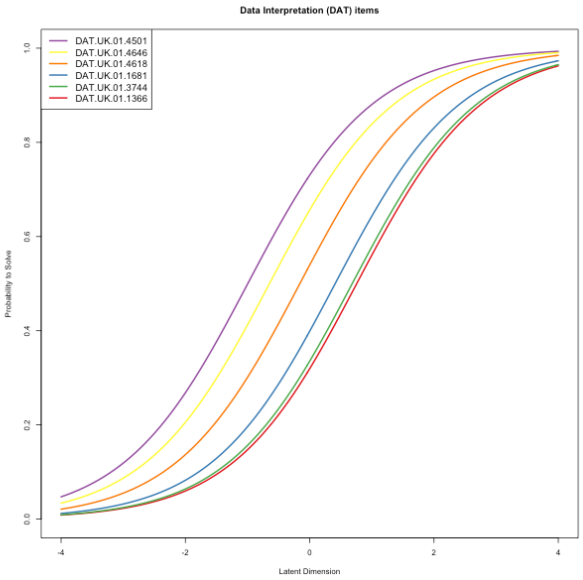




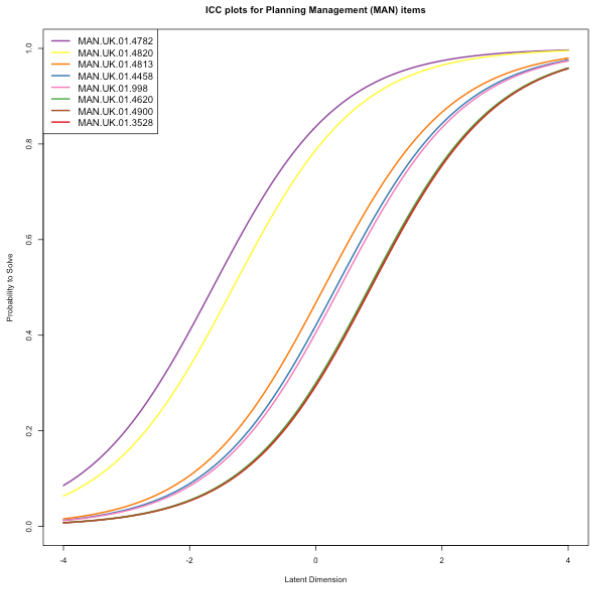


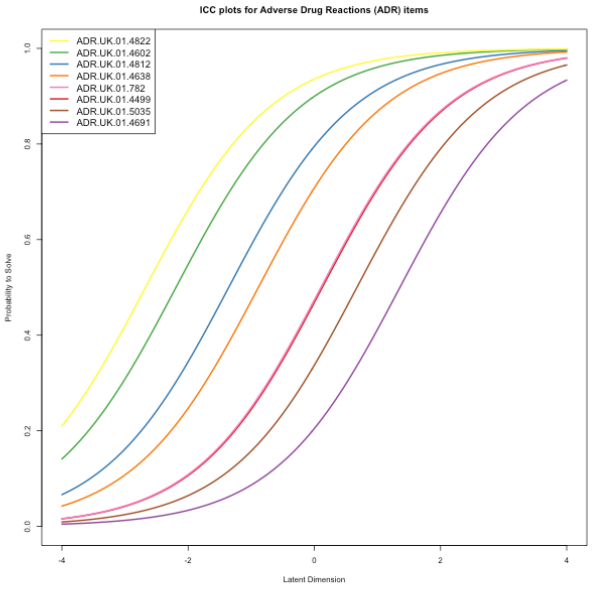


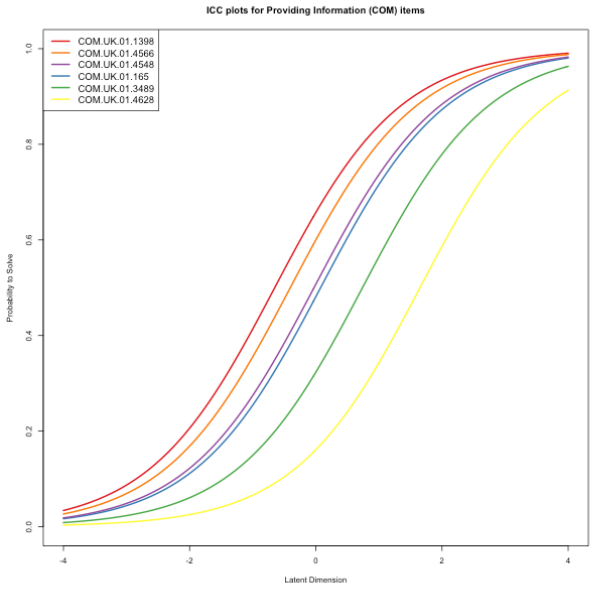


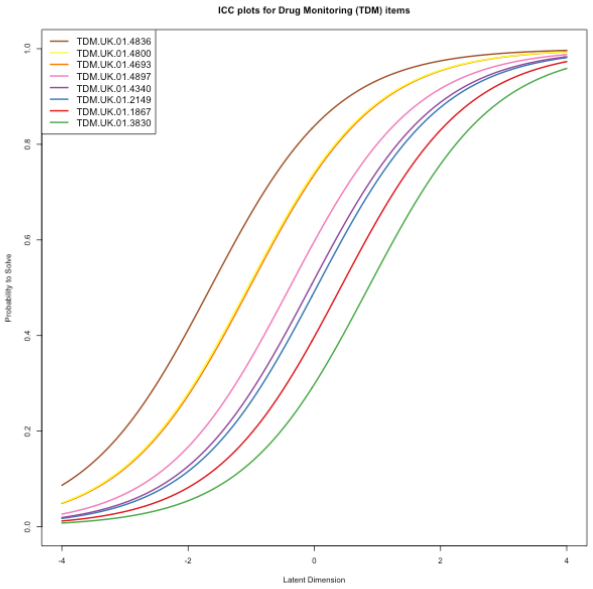


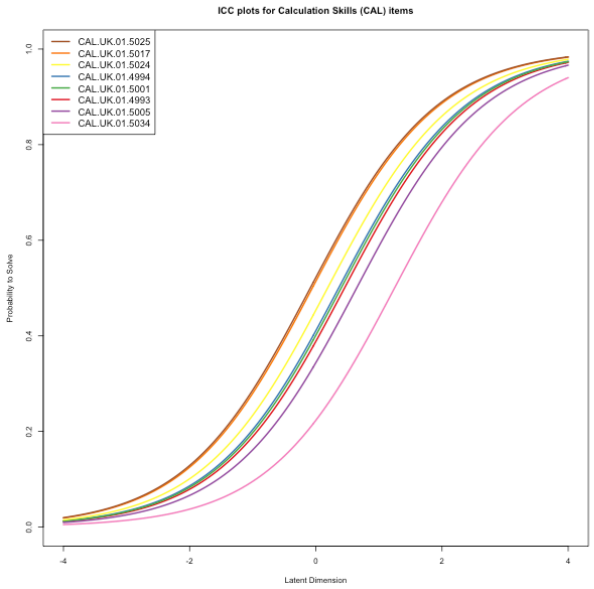
Paper C – RASCH model ICCs

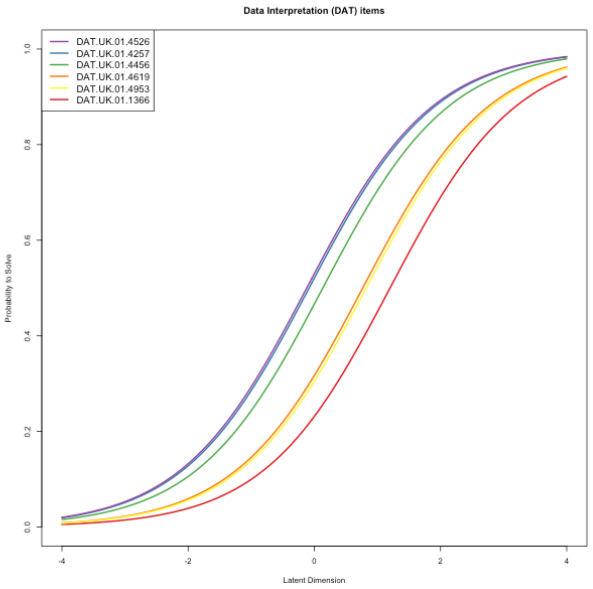




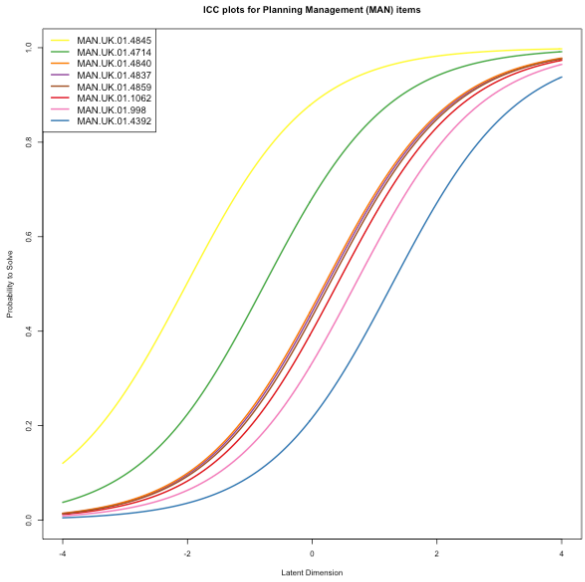


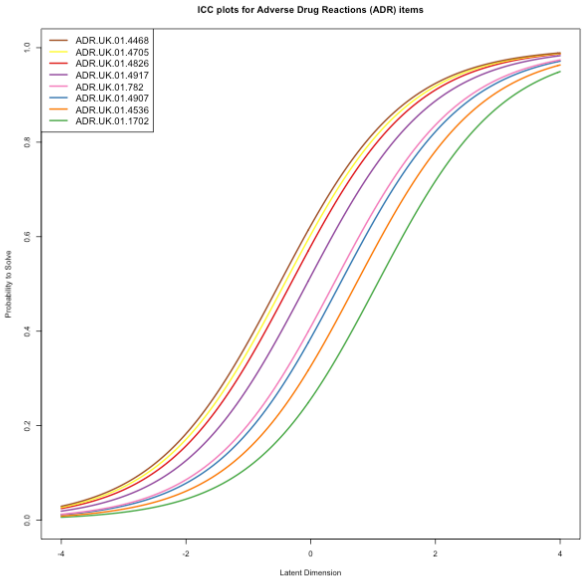


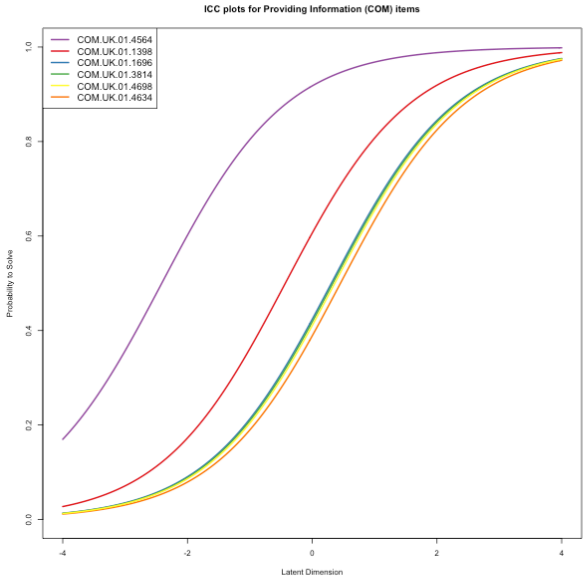


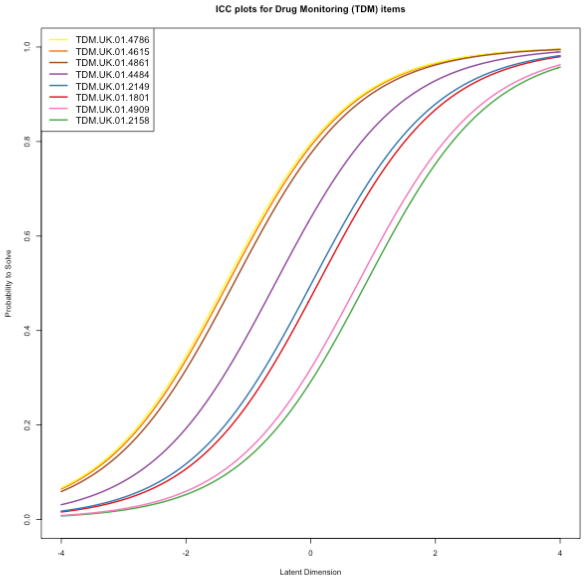


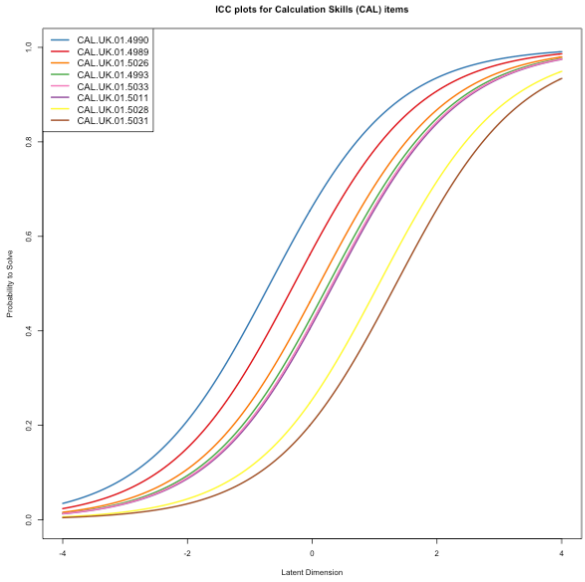
Paper D – RASCH model ICCs

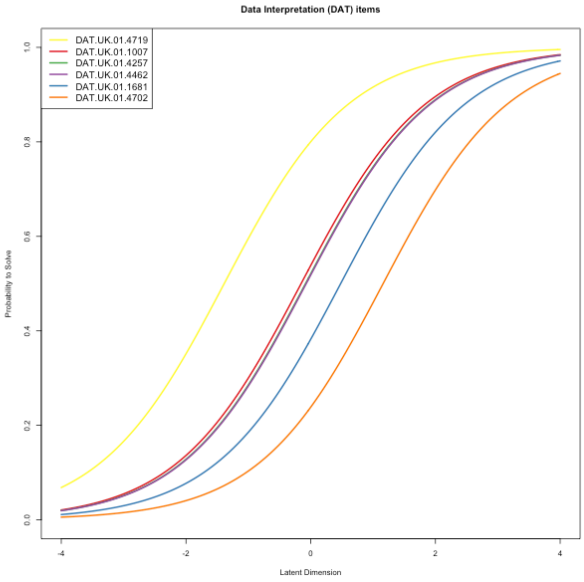






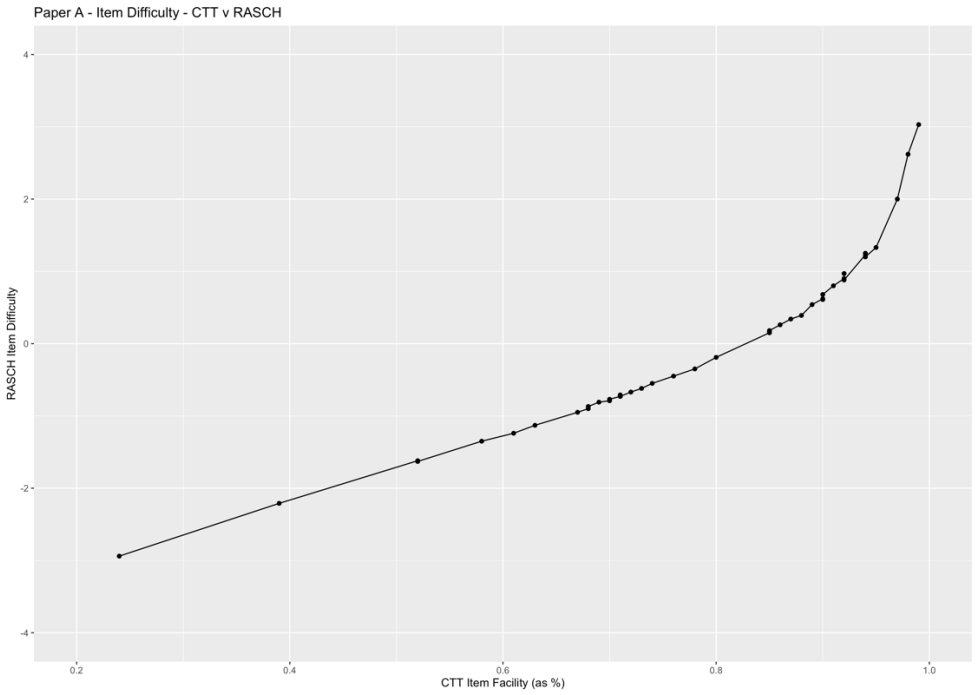


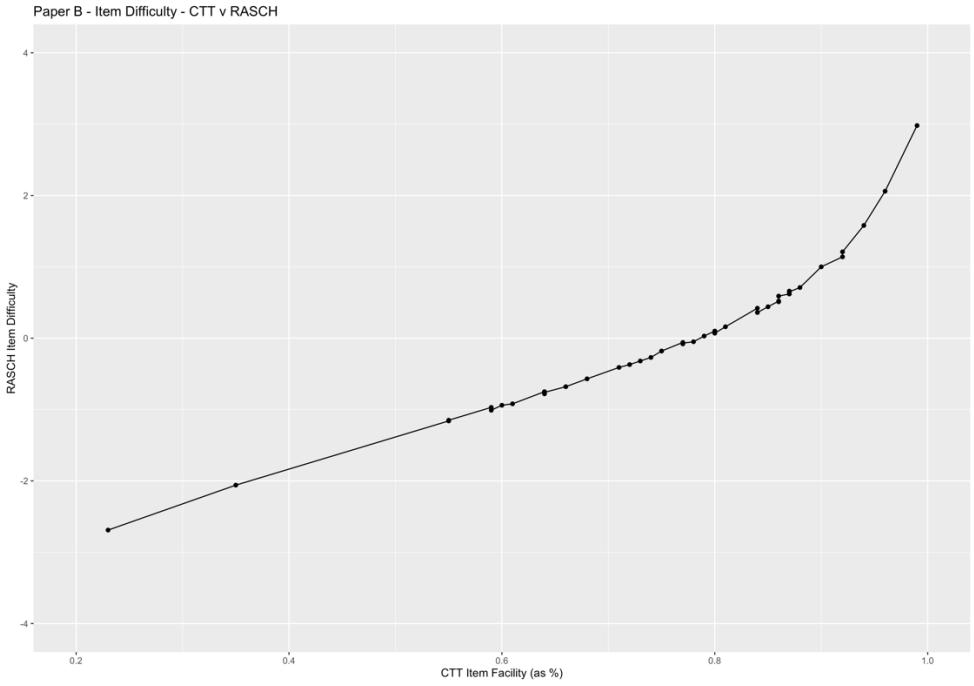


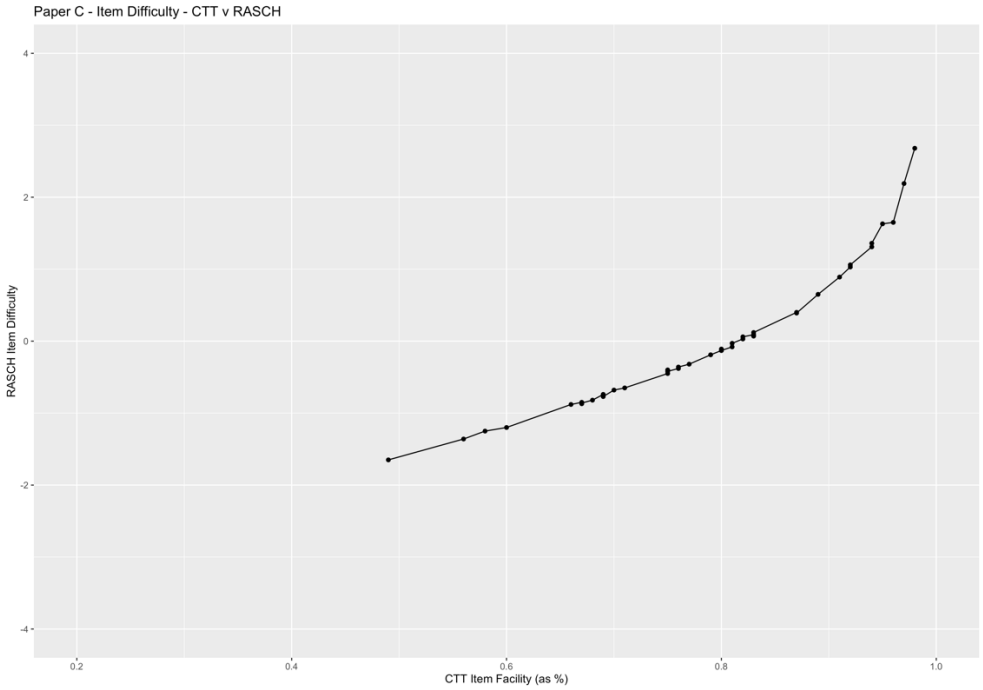


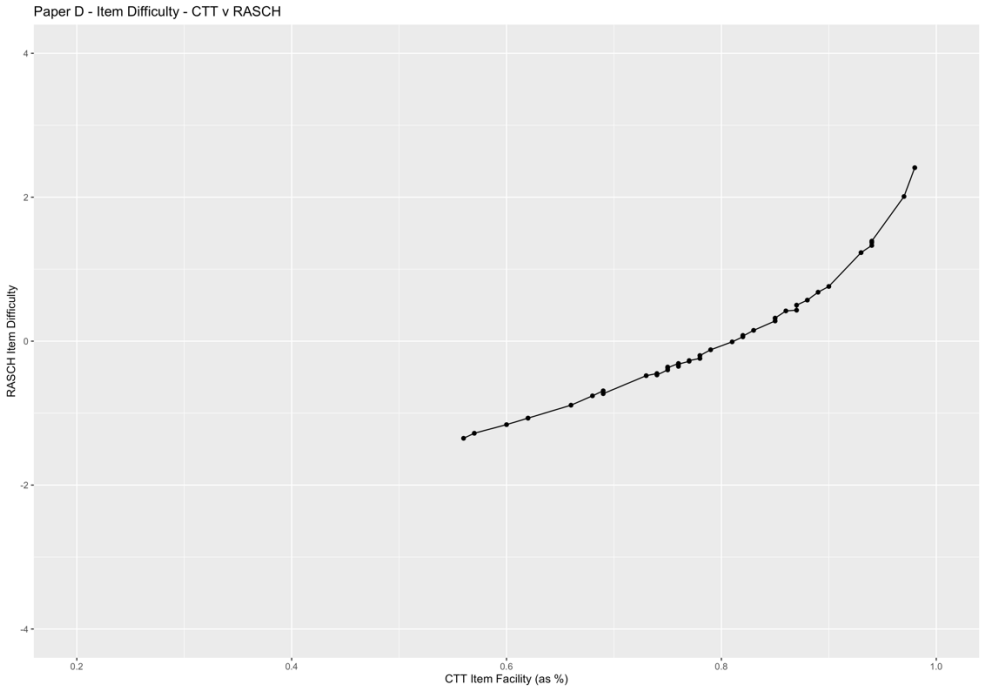
Comparison between CTT Facility and RASCH Beta item difficulty estimates

These plots compare the Item Facility from the classical test model to the RASCH beta difficulty parameters for the SBA items common to both.






Paper C - Item Difficulty - CTT v RASCH
RASCH Item Difficulty
4
2
0
-2
-4
0.2
0.4
0.6
0.8
1.0
CTT Item Facility (as %)


Paper D - Item Difficulty - CTT v RASCH
RASCH Item Difficulty
4
2
0
-2
-4
0.2
0.4
0.6
0.8
1.0
CTT Item Facility (as %)