University Clinical Aptitude Test (UCAT) Consortium
UCAT Examination

Executive Summary
Testing Interval: 1 August 2020 – 25 October 2020

Prepared by:
Pearson VUE
March 2021

Non-disclosure and Confidentiality Notice

This document contains confidential information concerning Pearson's services, products, data security procedures, data storage parameters, and data retrieval processes. You are permitted to view and retain this document provided that you disclose no part of the information contained herein to any outside agent or employee, except those agents and employees directly charged with reviewing this information. These agents and employees should be instructed and agree not to disclose this information for any purposes beyond the terms stipulated in the agreement of your company or agency with Pearson.

Copyright © 2021 NCS Pearson, Inc. All rights reserved. The PEARSON logo is a trademark in the U.S. and/or other countries.

Table of Contents

List of Tables

List of Figures

Introduction

The University Clinical Aptitude Test (UCAT) was administered in 2020 from 1 August to 25 October. During this period, a total of 34,144 exams were administered. Each exam consisted of four scored cognitive subtests: Verbal Reasoning (VR), Quantitative Reasoning (QR), Abstract Reasoning (AR), and Decision Making (DM). The cognitive subtests were followed by a Situational Judgement Test (SJT).

Candidate total cognitive score was higher than in previous years, and this was primarily caused by an increase in the scores on AR and DM. The mean scores on these subtests were higher by 15 points and seven points respectively. The mean QR and VR scores were higher by two points and five points respectively. The mean scores for the SJT also increased by 15 points, which had a corresponding increase on the band distribution, resulting in more candidates falling into Band 1 and fewer into Band 2, Band 3 and Band 4 than in previous years.

In terms of candidate performance by socio-economic classification (SEC) (UK candidates only; based on parental profession), Class 1 (Managerial and Professional Occupations) was consistently associated with higher mean scaled scores in the cognitive subtests. Class 4 (Lower Supervisory and Technical Occupations) and Class 5 (Semi-routine or Routine Occupations) were associated with the lowest mean scores. These trends are consistent with the performance by socio-economic groups observed in previous years.

Candidate age was broken down into five groups: ≤15, 16 to 19, 20 to 24, 25 to 34, and ≥35, and performance by age group was analysed in reference to candidates' highest educational qualifications. For candidates with Honours degrees, the age group 20 to 24 showed the highest scores across all cognitive sections. For candidates with school-leaving qualifications (i.e., below Honours degrees), the age group 16 to 19 had the highest scores. This pattern of performance is the same as that observed in 2019.

This report also includes analysis of performance by candidate first language (English vs. non-English for UK and non-UK residents) and country of residence. Consistent with 2019, the results indicate that candidates who reported English as their first language performed better on all cognitive sections than candidates who did not list English as their first language, and candidates who stated that they resided in the UK outperformed candidates who stated that they did not reside in the UK. The SJT showed similar trends to the cognitive sections by first language.

In 2020, the exam was delivered in two modes: standard test centre and online proctored. The online proctored mode allowed candidates to take the test without attending a test centre. Thirty-two percent of candidates took the exam in the online mode and 68% took the exam via the normal test centre route. Average test centre scores were rather higher than scores for candidates who took the test in the online mode. This was particularly notable for QR and AR, which were 16 and 18 points higher on average in the test centre than online. Further analysis indicates that these differences were primarily driven by demographic differences between the online and test centre cohorts.

Background

The UCAT Consortium was formed by various medical and dental schools of higher-education institutions in the United Kingdom. The purpose of the UCAT examination is to help select and/or identify more accurately those individuals with the innate ability to develop professional skills and competencies required to be a good clinician. The test results are to be used by institutions of higher education as part of the process of determining which applicants are to be accepted into the courses for which they have applied. The test results are also used by the Consortium for research to improve educational services. The goals of the Consortium are to use the UCAT to widen access for students who desire to study Medicine and Dentistry at university level and to admit those candidates who will become the very best doctors and dentists of the future.

The UCAT examination was first administered in July 2006 through the Pearson VUE Test Delivery System in testing centres in the United Kingdom and other countries. The 2020 testing period began on 1 August 2020 and ended on 25 October 2020. During this period, a total of 34,144 exams were administered. Five forms each of the VR, QR, AR, DM, and SJT subtests were used to generate five UCAT forms (Table 1). Each candidate was randomly assigned one of the five operational (scored) versions of the subtests and a set of pretest (unscored) items.

Table 1. Composition of the Five UCAT Forms

UCAT Form	VR	QR	AR	DM	SJT
Form 1	VR1	QR1	AR1	DM1	SJT1
Form 2	VR2	QR2	AR2	DM2	SJT2
Form 3	VR3	QR3	AR3	DM3	SJT3
Form 4	VR4	QR4	AR4	DM4	SJT4
Form 5	VR5	QR5	AR5	DM5	SJT5

The cognitive test forms were developed from the operational items used in the 2006 to 2019 administrations and from items that had been pretested during these years. The SJT items were developed from operational and pretest items used from 2013 to 2019. All items (operational and pretest) were analysed, and those with acceptable item statistics were saved as the active item bank.

Exam Design

The UCAT is an aptitude exam and is designed to measure innate cognitive abilities in addition to individuals' judgement regarding situations encountered in a target role. It is not an exam that measures student achievement, and therefore it does not contain any curriculum or science content.

The 2020 exam contained one SJT subtest and four cognitive subtests: VR, QR, AR and DM. Each exam was composed of 164 items on the cognitive subtests, of which 148 were operational items and 16 were pretest items. In addition, there were 69 SJT items, of which 63 were operational and six were pretest.

Candidates were given 120 minutes to answer a total of 233 items from the five subtests. There were four groups of candidates with time accommodation in 2020. Candidates with special educational needs (SEN) were allotted 150 minutes (UCATSEN) or 180 minutes (UCATSEN50) based on UCAT's pre-approval, and candidates with special accommodation (UCATSA) were allotted 120 minutes for the entire exam with flexible breaks, or 180 minutes for the entire exam with flexible breaks (UCATSENSA). The design of the exam is shown in Table 2.

Table 2. UCAT Exam Design

Subtest	Scored Items	Unscored Items	Total Number of Items	Test Time
VR	10 testlets of 4 items	1 testlet of 4 items	44	21 minutes allowed on items and 1 minute for instruction
QR	8 testlets of 4 items	1 testlet of 4 items	36	24 minutes allowed on items and 1 minute for instruction
AR	10 testlets of 5 items	1 testlet of 5 items	55	13 minutes allowed on items and 1 minute for instruction
DM	1 testlet of 26 items	3 items	29	31 minutes allowed on items and 1 minute for instruction
SJT	20 testlets of 1 to 5 items	1 testlet of 5 items 1 testlet of 1 item	69	26 minutes allowed on items and 1 minute for instruction

Examination Results

Cognitive Subtests

Candidates' scaled scores are reported for each of the four cognitive subtests and are based on all scored items in each subtest. The cognitive subtest scaled scores range from 300 to 900. Universities receive the subtest scaled scores for each student plus a total score that is a simple sum of the four subtest scores and has a range of 1,200 to 3,600. An item response theory (IRT) calibration model and IRT true-score equating methods were used to transform the raw scores from each form into a common reporting scale.

Table 3 presents summary statistics for each of the cognitive subtests plus the total summed scaled score for the total group. The scaled score means vary across the four cognitive subtests. The mean scaled score for all candidates was 570 for VR, 664 for QR, 653 for AR, and 625 for DM. Standard deviations ranged from 74 (VR) to 93 (AR).

Table 3. Cognitive Subtest and Total Scaled Score Summary Statistics: Total Group

Test	Total *N*	Mean	*SD*	Min	Max
VR	34,144	569.76	74.25	300	900
QR	34,144	664	78.2	300	900
AR	34,144	652.78	93.11	300	900
DM	34,144	624.89	88.79	300	900
Total	34,144	2,511.43	269.16	1,410	3,490

Table 4 summarises the scaled score statistics for UCAT non-SEN candidates and SEN candidates. SEN candidates outperformed non-SEN candidates in all four subtests. However, the sample sizes of UCATSEN50, UCATSA and UCATSENSA are small, and the results should be treated with caution.

Table 4. Cognitive Subtest and Total Scaled Score Summary Statistics: SEN vs. non-SEN

Exam	Test	Total *N*	Mean	*SD*	Min	Max
UCAT	VR	32,297	568.66	73.95	300	900
	QR	32,297	662.74	77.92	330	900
	AR	32,297	651.63	92.82	300	900
	DM	32,297	623.78	88.79	300	900
	Total	32,297	2,506.82	268.73	1,420	3,490
UCATSEN	VR	1,501	587.02	76.49	350	900
	QR	1,501	683.28	80.11	300	900
	AR	1,501	672.97	96.43	300	900
	DM	1,501	640.26	86.96	300	900
	Total	1,501	2,583.53	265.6	1,410	3,310
UCATSENSA	VR	154	603.83	76.48	350	820
	QR	154	705.06	72.37	530	900
	AR	154	674.29	92.02	360	890

Exam	Test	Total *N*	Mean	*SD*	Min	Max
	DM	154	680.06	81.33	460	880
	Total	154	2,663.25	248.32	2,010	3,440
UCATSEN50	VR	137	596.57	82.33	320	870
	QR	137	698.83	87.06	500	880
	AR	137	674.96	96.04	300	890
	DM	137	641.17	86.02	360	880
	Total	137	2,611.53	272.66	1,610	3,250
UCATSA	VR	55	580.55	72.17	350	740
	QR	55	676.18	62.82	550	830
	AR	55	660.91	89.7	460	880
	DM	55	661.09	66.07	530	800
	Total	55	2,578.73	207.03	2,140	2,970

Situational Judgement Test

For the SJT, candidates are awarded one of four bands to reflect their performance on the operational items in the test. The bands are determined using the scaled score calculated for each candidate, as shown in Table 5.

The scaled score, which is not issued to candidates, ranges from 300 to 900. The scaled score is designed to place proportions of candidates into each band based on the 2018 score distribution.

A classical pre-equating model was used to transform the raw scores from each form onto a common reporting scale. As the psychometric model used for the SJT is different to that used for the cognitive subtests, the scores are not directly comparable.

Table 5. SJT Band Scaled Score Range and Description (Base in 2018)

Bands	Scaled Score Range	Intended Band Proportions	Narrative
Band 1	655-900	22%	Those in Band 1 demonstrated an excellent level of performance, showing similar judgement in most cases to the panel of experts.
Band 2	594-654	38%	Those in Band 2 demonstrated a good, solid level of performance, showing appropriate judgement frequently, with many responses matching model answers.
Band 3	518-593	30%	Those in Band 3 demonstrated a modest level of performance, with appropriate judgement shown for some questions and substantial differences from ideal responses for others.
Band 4	300-517	10%	The performance of those in Band 4 was low, with judgement tending to differ substantially from ideal responses in many cases.

Table 6 presents the number and percentage of candidates in each band for the 34,144 candidates who took the UCAT during the 2020 testing window. The proportions observed in the 2020 SJT differed somewhat from the intended percentages. Although Band 4 was only one point away from the intended target, Band 1 was eight points higher at the expense of Band 2 and Band 3, which were two points and six points lower than intended respectively.

Table 6. SJT Band Distribution in 2020

SJT Band	Number of Candidates	Percentage of Candidates
Band 1	10,404	30%
Band 2	12,455	36%
Band 3	8,064	24%
Band 4	3,221	9%
Total	34,144	100%

Table 7 summarises the percentage by band and the scaled score statistics for SEN and non-SEN candidates. SEN candidates outperformed non-SEN candidates on the SJT.

Table 7. SJT Percentage by Band and Summary Statistics for SEN and non-SEN Candidates

Exam	Total *N*	Percentage of Candidates				Scaled Score			
		Band 1	Band 2	Band 3	Band 4	Mean	*SD*	Min	Max
UCAT	32,297	30%	36%	24%	10%	612.35	70.03	300	776
UCATSEN	1,501	37%	39%	19%	5%	628.45	61.3	300	764
UCATSENSA	154	47%	35%	15%	3%	642.83	57.57	449	768
UCATSEN50	137	42%	34%	18%	5%	629.94	69.8	300	744
UCATSA	55	40%	38%	20%	2%	637.69	49.17	508	719

Examination Results by Demographic Variables

For the purpose of the demographic analysis, the SJT scaled score summary statistics are included in the relevant tables to illustrate trends. These scores are not issued to candidates and are not directly comparable to the cognitive subtest scaled scores.

Gender

Table 8 presents scaled score summary statistics for males and females for each of the subtests. Females constituted 21,761 (64%) candidates and males 12,324 (36%). On average, males outperformed females on cognitive subtests. Female candidates outperformed male candidates on the SJT. The proportion of female to male candidates is the same as in 2019 and the pattern of males outperforming females in the cognitive sections and vice versa in the SJT is also consistent with previous years.

Table 8. Subtest and Total Scaled Score Summary Statistics by Gender

Test	Gender	Total *N*	Total %	Mean	*SD*	Min	Max
VR[1]	Female	21,761	64%	566.2	73.57	300	890
	Male	12,324	36%	575.9	75.03	300	900
QR[2]	Female	21,761	64%	654.63	75.38	300	900
	Male	12,324	36%	680.47	80.3	390	900
AR[3]	Female	21,761	64%	650.24	92.01	300	900
	Male	12,324	36%	657.23	94.8	300	900
DM[4]	Female	21,761	64%	619.77	88.75	300	900
	Male	12,324	36%	633.74	88.09	300	900
Total Cognitive Scaled Score[5]	Female	21,761	64%	2,490.85	266.32	1,410	3,490
	Male	12,324	36%	2,547.33	270.17	1,460	3,490
SJT[6]	Female	21,761	64%	618.67	67.6	300	776
	Male	12,324	36%	603.76	72.24	300	770

Ethnicity

Table 9 summarises the performance of the various ethnic groups on each of the four cognitive subtests. Only UK candidates are asked to provide an ethnic group. The categories have been collated as follows:

- UK–White: White
- UK–Asian: Asian Indian; Asian Pakistani; Asian Bangladeshi; Asian Other
- UK–Black: Black Caribbean; Black African; Black Other
- UK–Mixed Race: Mixed White and Black Caribbean; Mixed White and Black African; Mixed White and Asian; Other Mixed

[1] T-statistics = 11.61 (df=34083, p<0.01) assuming equal variance.

[2] T-statistics = 29.69 (df=34083, p<0.01) assuming equal variance.

[3] T-statistics = 6.66 (df=34083, p<0.01) assuming equal variance.

[4] T-statistics = 13.99 (df=34083, p<0.01) assuming equal variance.

[5] T-statistics = 18.71 (df=34083, p<0.01) assuming equal variance.

[6] T-statistics = -19.09 (df=34083, p<0.01) assuming equal variance.

- UK–Chinese: Asian Chinese
- UK–Other: Other, e.g. gypsy, traveller, or Irish traveller, or not specified.

Table 9. Subtest and Total Scaled Score Summary Statistics by Ethnic Group

Test	Ethnic Group	Total *N*	Total %	Mean	*SD*	Min	Max
VR[7]	Non-UK	5,759	17%	556	77	300	890
	UK–Asian	10,779	32%	559	68	300	870
	UK–Black	3,029	9%	548	66	300	870
	UK–Chinese	429	1%	587	77	380	870
	UK–Mixed Race	1,518	4%	578	77	320	890
	UK–Other	2,556	8%	560	76	300	890
	UK–White	9,877	29%	597	72	320	900
QR[8]	Non-UK	5,759	17%	656	84	330	900
	UK–Asian	10,779	32%	663	77	330	900
	UK–Black	3,029	9%	627	72	300	900
	UK–Chinese	429	1%	706	86	450	900
	UK–Mixed Race	1,518	4%	667	76	430	900
	UK–Other	2,556	8%	652	76	390	900
	UK–White	9,877	29%	682	72	390	900
AR[9]	Non-UK	5,759	17%	638	97	300	900
	UK–Asian	10,779	32%	656	93	300	900
	UK–Black	3,029	9%	618	87	300	900
	UK–Chinese	429	1%	708	101	440	900
	UK–Mixed Race	1,518	4%	658	97	300	900
	UK–Other	2,556	8%	646	87	300	900
	UK–White	9,877	29%	668	88	300	900
DM[10]	Non-UK	5,759	17%	616	92	300	890
	UK–Asian	10,779	32%	610	85	300	890
	UK–Black	3,029	9%	586	82	300	880
	UK–Chinese	429	1%	653	84	370	880
	UK–Mixed Race	1,518	4%	634	90	330	900
	UK–Other	2,556	8%	613	92	300	900
	UK–White	9,877	29%	659	80	330	900
Total Cognitive Scaled Score[11]	Non-UK	5,759	17%	2,466	283	1,440	3,420
	UK–Asian	10,779	32%	2,488	261	1,420	3,490
	UK–Black	3,029	9%	2,378	245	1,410	3,400
	UK–Chinese	429	1%	2,655	280	1,880	3,300

[7] F-statistics = 374.2 (df=6, p<0.01).

[8] F-statistics = 262 (df=6, p<0.01).

[9] F-statistics = 173.8 (df=6, p<0.01).

[10] F-statistics = 454.1 (df=6, p<0.01).

[11] F-statistics = 433.1 (df=6, p<0.01).

Test	Ethnic Group	Total *N*	Total %	Mean	*SD*	Min	Max
	UK–Mixed Race	1,518	4%	2,537	274	1,610	3,490
	UK–Other	2,556	8%	2,471	267	1,620	3,280
	UK–White	9,877	29%	2,606	241	1,730	3,450
SJT[12]	Non-UK	5,759	17%	588	79	300	770
	UK–Asian	10,779	32%	613	68	300	768
	UK–Black	3,029	9%	604	72	300	752
	UK–Chinese	429	1%	628	58	440	759
	UK–Mixed Race	1,518	4%	620	66	300	751
	UK–Other	2,556	8%	609	74	300	759
	UK–White	9,877	29%	631	58	300	776

The UK–White ethnic group made up 29% of the testing population. Proportions for the other ethnic groups ranged from 1% to 32%. There was considerable variation in means among the different ethnic groups. For VR, DM and the SJT, the highest-performing group was UK–White. For QR and AR, the highest-performing group was UK–Chinese. This is consistent with scores patterns that were observed in the 2019 exam.

Socio-Economic Classification

Table 10 provides scaled score summary statistics for all UK candidates by SEC. For all cognitive subtests the means generally trended downwards in order of the occupational classes, from Class 1 to Class 5. For the SJT, Class 1 had the highest mean scaled score and Class 4 had the lowest. In other words, candidates whose parents/guardians had managerial or professional occupations tended to outperform those whose parents had lower supervisory, technical, or routine occupations.

Table 10. Subtest and Total Scaled Score Summary Statistics by NS-SEC Class for UK Candidates

Test	NS-SEC Class	Total *N*	Total %	Mean	*SD*	Min	Max
VR[13]	1	18,164	64%	581.73	73.31	320	900
	2	1,219	4%	576.46	71.11	320	820
	3	1,810	6%	556.47	66.47	320	900
	4	779	3%	547.55	67.33	300	790
	5	1,872	7%	554.41	67.82	300	890
	NA	4,541	16%	552.39	73.17	300	890
QR[14]	1	18,164	64%	674.35	76.40	300	900
	2	1,219	4%	664.34	71.56	390	880
	3	1,810	6%	653.01	73.40	390	900
	4	779	3%	647.02	75.20	430	900
	5	1,872	7%	649.25	73.70	330	900
	NA	4,541	16%	646.31	77.43	330	900
	1	18,164	64%	664.21	90.87	300	900

[12] F-statistics = 270.1 (df=6, p<0.01).

[13] F-statistics =230.6 (df=5, p<0.01).

[14] F-statistics = 146.17 (df=5, p<0.01).

Test	NS-SEC Class	Total *N*	Total %	Mean	*SD*	Min	Max
AR[15]	2	1,219	4%	653.27	88.72	330	900
	3	1,810	6%	643.14	90.40	330	900
	4	779	3%	639.35	90.67	360	900
	5	1,872	7%	643.28	92.82	300	900
	NA	4,541	16%	636.05	93.50	300	900
DM[16]	1	18,164	64%	640.13	86.71	300	900
	2	1,219	4%	627.54	80.32	320	880
	3	1,810	6%	603.55	84.30	330	880
	4	779	3%	595.67	84.31	330	880
	5	1,872	7%	600.40	81.24	300	870
	NA	4,541	16%	597.97	87.94	300	880
Total Cognitive Scaled Score[17]	1	18,164	64%	2560.42	260.16	1,410	3,490
	2	1,219	4%	2521.62	243.32	1,600	3,350
	3	1,810	6%	2456.17	250.35	1,740	3,400
	4	779	3%	2429.59	252.57	1,620	3,220
	5	1,872	7%	2447.34	250.52	1,510	,320
	NA	4,541	16%	2432.72	269.49	1,420	,490
SJT[18]	1	18,164	64%	623.78	63.09	300	776
	2	1,219	4%	623.01	62.41	300	755
	3	1,810	6%	610.22	69.21	319	760
	4	779	3%	608.34	67.90	317	752
	5	1,872	7%	610.98	67.53	326	749
	NA	4,541	16%	604.31	75.01	300	759

Note. Codes for SEC Categories
1 – Managerial and Professional Occupations
2 – Intermediate Occupations
3 – Small Employers and Own Account Workers
4 – Lower Supervisory and Technical Occupations
5 – Semi-routine and Routine Occupations
NA – Could not calculate SEC group, i.e. information withheld

Age and Education

Table 11 provides scaled score summary statistics for the total group both by age group and the candidates' highest educational qualification. Candidates were divided into five age groups: ≤15, 16 to 19, 20 to 24, 25 to 34, and ≥35. Two categories of educational qualification were examined: Below Honours Degree and Honours Degree or Above. Candidates in the Honours Degree or Above category were mostly in the 20 to 24 age group, which also represented the highest mean scores across all four cognitive subtests for that category of educational qualification. Candidates in the Below Honours Degree category were mostly in the 16 to 19 age group, which showed the highest mean scores across all four cognitive subtests for that category of educational qualification.

[15] F-statistics = 127.4 (df=5, p<0.01).

[16] F-statistics = 256.24 (df=5, p<0.01).

[17] F-statistics = 285.56 (df=5, p<0.01).

[18] F-statistics = 238.26 (df=5, p<0.01).

Table 11. Subtest and Total Scaled Score Summary Statistics by Age Group and Highest Qualification

Test	Highest Qualification	Age Group	Total *N*	% Total *N*	Mean	*SD*	Min	Max
VR	Below Honours Degree	Up to 15	58	0%	549.483	82.2368	300	870
		16-19	25,578	96%	569.52	72.84	300	900
		20-24	723	3%	540.91	82.40	300	820
		25-34	192	1%	540.83	85.89	300	790
		>=35	46	0%	539.78	81.69	390	740
	Honours Degree or Above	Up to 15	0	0%	NA	NA	NA	NA
		16-19	369	5%	540.00	72.12	370	870
		20-24	5,255	70%	578.23	73.98	300	900
		25-34	1,666	22%	575.26	81.61	300	890
		>=35	257	3%	539.46	87.62	320	820
QR	Below Honours Degree	Up to 15	58	0%	626.724	79.7229	460	900
		16-19	25,578	96%	668.39	77.88	300	900
		20-24	723	3%	637.63	92.29	430	900
		25-34	192	1%	611.98	77.94	390	880
		>=35	46	0%	592.39	62.33	480	800
	Honours Degree or Above	Up to 15	0	0%	NA	NA	NA	NA
		16-19	369	5%	631.30	77.52	450	900
		20-24	5,255	70%	661.94	74.20	330	900
		25-34	1,666	22%	641.22	74.69	330	900
		>=35	257	3%	598.72	67.98	430	800
AR	Below Honours Degree	Up to 15	58	0%	611.724	86.5952	460	880
		16-19	25,578	96%	655.59	92.26	300	900
		20-24	723	3%	622.23	101.09	300	900
		25-34	192	1%	588.91	91.38	300	900
		>=35	46	0%	578.91	91.95	360	800
	Honours Degree or Above	Up to 15	0	0%	NA	NA	NA	NA
		16-19	369	5%	623.04	86.24	400	880
		20-24	5,255	70%	658.75	92.24	300	900
		25-34	1,666	22%	632.34	92.44	300	900
		>=35	257	3%	581.79	94.70	300	880
DM	Below Honours Degree	Up to 15	58	0%	591.035	91.7412	420	830
		16-19	25,578	96%	629.02	87.64	300	900
		20-24	723	3%	585.66	103.83	300	880
		25-34	192	1%	565.05	97.46	360	870
		>=35	46	0%	541.52	98.79	330	800
	Honours Degree or Above	Up to 15	0	0%	NA	NA	NA	NA
		16-19	369	5%	580.70	91.40	300	830
		20-24	5,255	70%	625.44	85.09	300	890
		25-34	1,666	22%	607.61	90.51	300	890

Test	Highest Qualification	Age Group	Total *N*	% Total *N*	Mean	*SD*	Min	Max
		>=35	257	3%	555.64	95.59	300	830
Total Score	Below Honours Degree	Up to 15	58	0%	2378.97	278.188	1900	3220
		16-19	25,578	96%	2,522.53	265.39	1,410	3,490
		20-24	723	3%	2,386.43	321.37	1,480	3,290
		25-34	192	1%	2,306.77	295.96	1,610	3,390
		>=35	46	0%	2,252.61	275.53	1,650	3,090
	Honours Degree or Above	Up to 15	0	0%	NA	NA	NA	NA
		16-19	369	5%	2,375.04	267.29	1,590	3,190
		20-24	5,255	70%	2,524.37	256.76	1,510	3,440
		25-34	1,666	22%	2,456.44	274.50	1,420	3,320
		>=35	257	3%	2,275.60	287.44	1,550	3,030
SJT	Below Honours Degree	Up to 15	58	0%	543.586	82.1872	300	695
		16-19	25,578	96%	609.72	68.93	300	776
		20-24	723	3%	593.60	82.21	302	755
		25-34	192	1%	585.90	92.32	324	745
		>=35	46	0%	590.93	81.20	412	734
	Honours Degree or Above	Up to 15	0	0%	NA	NA	NA	NA
		16-19	369	5%	581.28	82.22	300	730
		20-24	5,255	70%	634.41	60.85	300	764
		25-34	1,666	22%	628.46	69.20	300	766
		>=35	257	3%	581.41	105.15	300	742

Similar to cognitive subtests, the Below Honours Degree category had the highest mean SJT scaled scores at ages 16 to 19, and the Honours Degree or Above category had the highest mean SJT score at ages 20 to 24. These trends are consistent with those observed in previous years.

First Language

Scaled score analysis by candidate first language (English vs. Other for UK and non-UK candidates) is presented in Table 12. Candidates whose first language is English performed better on all subtests, including the SJT, compared to candidates whose first language is not English. Similarly, candidates whose residency was UK outperformed candidates whose residency was not the UK in all subtests including the SJT.

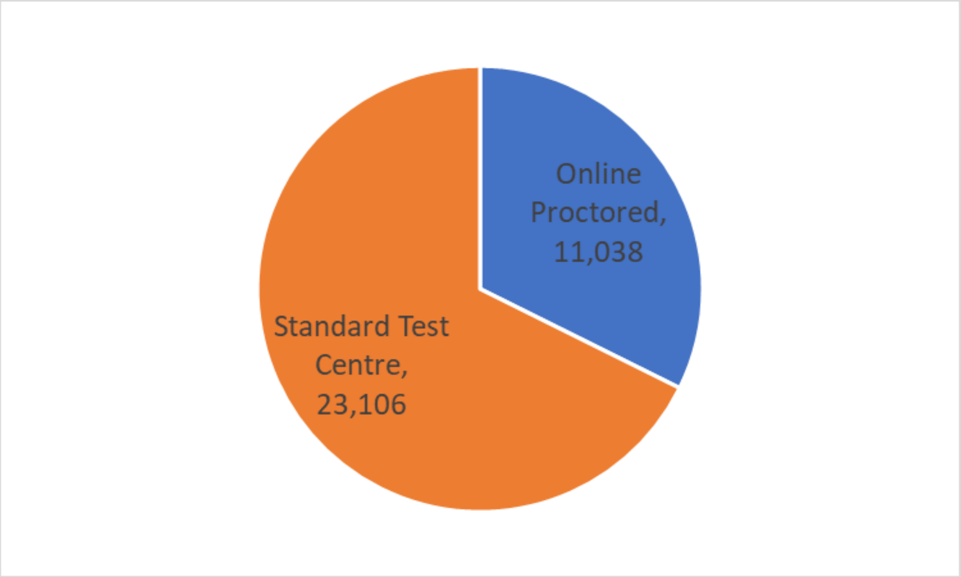
Table 12. Subtest and Total Scaled Score Summary Statistics by Country of Residence and First Language

Test	Country of Residence	First Language	Total *N*	% Total *N*	Mean	*SD*	Min	Max
VR	Non-UK	English	2,641	46%	578.31	75.30	300	890
		Other	3,118	54%	537.92	72.89	300	870
	UK	English	20,160	71%	583.66	72.89	300	900
		Other	8,225	29%	545.00	67.36	300	890
QR	Non-UK	English	2,641	46%	665.63	84.35	390	900
		Other	3,118	54%	647.47	82.39	330	900
	UK	English	20,160	71%	673.57	75.61	300	900
		Other	8,225	29%	646.30	76.66	330	900
AR	Non-UK	English	2,641	46%	642.35	94.53	300	900
		Other	3,118	54%	633.82	98.56	300	900
	UK	English	20,160	71%	661.19	91.11	300	900
		Other	8,225	29%	642.69	92.97	300	900
DM	Non-UK	English	2,641	46%	634.90	89.16	330	890
		Other	3,118	54%	600.20	90.64	300	890
	UK	English	20,160	71%	640.50	85.25	300	900
		Other	8,225	29%	592.78	85.78	300	890
Total Score	Non-UK	English	2,641	46%	2521.20	275.71	1540	3420
		Other	3,118	54%	2419.41	280.00	1440	3390
	UK	English	20,160	71%	2558.92	257.27	1410	3490
		Other	8,225	29%	2426.77	261.68	1420	3490
SJT	Non-UK	English	2,641	46%	602.95	70.33	300	770
		Other	3,118	54%	574.80	84.07	300	752
	UK	English	20,160	71%	625.05	62.05	300	776
		Other	8,225	29%	602.43	73.57	300	762

Comparison of Examination Results by Mode

In previous years, the UCAT exam has been delivered only in test centres. In 2020, the exam was also offered in the online proctored mode. As illustrated in Figure 1, 32% of candidates opted for the online mode, and 68% of candidates opted for the standard test centre mode.

Figure 1. Distribution of Candidates by Mode of Delivery



Candidates who sat the exam in the test centre mode outperformed the online candidates, as shown in Table 13. This was most pronounced in AR, where, on average, test centre candidates achieved 18 more scaled score points than the online candidates. It was least pronounced in VR, where the difference was only four points. The mean total cognitive score difference between the modes was 48 points.

Table 13. Mean Subtest Scaled Score by Mode of Delivery

	Mean Scaled Score	
Subtest	Online Proctored	Standard Test Centre
VR	567	571
QR	653	669
AR	641	659
DM	619	628
Total	2,479	2,527

Table 14 shows candidate demographics by mode of delivery. The online proctored mode had proportionally fewer UK candidates and more Honours degree candidates. This may explain why candidates in the online mode underperformed relative to candidates in the standard test centre mode.

Table 14. Candidate Demographic by Mode of Delivery

		Online Proctored		Standard Test Centre	
	Demographic	*N* Candidates	% Candidates	*N* Candidates	% Candidates
Permanent Residence	Non-UK	2,683	0.24	3,076	0.13
	UK	8,363	0.76	20,022	0.87
English Fluency	No	3,762	0.34	7,581	0.33
	Yes	7,284	0.66	15,517	0.67
Gender	Female	6,981	0.63	14,780	0.64
	Male	4,044	0.37	8,280	0.36
Highest Qualification	Below Honours Degree	8,228	0.74	18,369	0.8
	Honours Degree or Above	2,818	0.26	4,729	0.2
Age	Up to 15	2	0	2	0
	16-19	7,984	0.72	17,963	0.78
	20-24	2,200	0.2	3,778	0.16
	25-34	736	0.07	1,122	0.05
	>=35	105	0.01	198	0.01

Consistent with the difference in scaled scores, for the SJT, the standard test centre candidates received Band 1 and Band 2 proportionally more, and Band 3 and Band 4 proportionally less, than the online candidates (Table 15).

Table 15. Candidate SJT Band by Mode of Delivery

	Online Proctored		Standard Test Centre		
SJT Band	*N* Candidates	% Candidates	*N* Candidates	% Candidates	Target
Band 1	3,036	28%	7368	32%	22%
Band 2	3881	35%	8574	37%	38%
Band 3	2793	25%	5271	23%	30%
Band 4	1328	12%	1,893	8%	10%
Total	11038	100%	23106	100%	

Pearson VUE employed two methods to examine the difference in scores between modes: stratified sampling; and item drift comparison.

Stratified Sampling

Stratified sampling is effective for examining differences in scores between two groups because it allows for a population to be sampled in reference to specific characteristics, in this case demographics. A sample of 10,000 candidates was taken from the standard test centre candidate population. This sample was selected in such a way that its demographic characteristics matched those of the online proctored candidate group.

Comparing the mean scores of the online candidates with those of one sample of test centre candidates ensures the demographic characteristics of the groups match; however, there may be influence by random variation, which is not controlled by using a single sample. In order to manage this variation, the sample was taken 100 times.

The mean scaled score for each subtest across all 100 stratified samples is presented in Table 19 below. Table 19 shows that when the demographic characteristics of the two modes are similar, the difference in mean scores by subtest falls considerably. The smallest fall was in VR, where the original difference between online and test centre scores was 4, which falls to 1 after the demographics are aligned. The largest difference was in AR, where the difference falls from 18 to 8.

Table 16. Score Before and After Sampling

Subtest	Online Proctored	Standard Test Centre before Sampling	Standard Test Centre after Sampling	Difference between Modes before Sampling	Difference between Modes after Sampling
VR	567	571	568	4	1
QR	653	669	660	16	7
AR	641	659	648	18	8
DM	619	628	621	9	2
Total	2,479	2,527	2,497	48	18

This analysis indicates that the score difference between modes is indeed a reflection of the demographic characteristics of the groups.

Item Drift

Item drift analysis is a method for examining changes in item difficulty over time. Each operational cognitive item has a fixed difficulty value that was defined when the item was pretested. After each exam period, the items are calibrated again, and if they diverge significantly from their fixed values, they are considered to have drifted.

In order to examine whether there is an item-level influence on candidate performance by mode, it is possible to calibrate each item by cohort and examine the difference between the number of items that drift. The items were calibrated using Rasch item response theory for three cohorts of candidates: 1) online candidates only; 2) test centre candidates only; and 3) online and test centre candidates combined. Table 17 shows the outcome of this analysis for the subtests that had the largest differences in mean scores between modes: QR and AR.

QR had no significant drift when the modes were isolated and one item that drifted positively when all candidate responses were analysed together. A logit is a unit of measurement used to reflect item difficulty in item response theory. Table 17 shows that the QR item drifted by 0.08 logits. The difficulty of the operational QR items ranged from -1.86 to 1.33 logits, so in its context, 0.08 logits is a small amount.

Overall, three AR items drifted. When separated out into the modes, this number did not change greatly. As Table 17 shows, three items drifted when only the online candidate data is used and four when only the test centre data is used. The average degree of drift is similarly small at 0.11 logits.

Table 17. Drift by Mode

	Number of Significant Drift			
Subtest	Online	Test centre	Online and Test Centre	Average Degree of Drift
QR	0	0	1	0.08 logits
AR	3	4	3	0.11 logits

Summary of Comparison of Results by Mode

Candidates who took the exam in test centres on average performed better than candidates who took the exam in the online mode. There are several demographic characteristics that are known to predict better performance, so Pearson VUE explored whether demographic differences explain the score differences.

Employing a stratified sampling approach to comparing average performance in each mode demonstrates that a considerable proportion of the score differences can be accounted for by differences in demographics. Additionally, the item drift analysis shows limited apparent variation in the difficulty of individual items between modes, meaning the difference in scores between modes does not appear to be related to individual items being more difficult in one mode than another.

Test and Item Analysis

Test analysis for the operational forms included computation of the raw score means, standard deviations, internal consistency reliabilities, and standard errors of measurement (*SEM*s) for each form of each cognitive subtest. Similar test analyses were performed and reported for the scaled scores for the cognitive subtests.

Item analysis for the cognitive subtests included a complete classical analysis of item characteristics including *p* values and point biserial (item discrimination). IRT analyses included an estimation of item difficulty, or the *b* parameter.

Test Analysis – Cognitive Subtests

The raw score means, standard deviations, ranges, internal consistency reliabilities (Cronbach's alpha), and *SEM*s for each form of each subtest are summarised in Table 18. The highest raw score reliabilities were found for AR. This can be attributed to the test length as AR has the largest number of items; generally, reliability increases with test length.

Table 18. Raw Score Test Statistics

Subtest	Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Alpha	*SEM*
VR	1	40	8,424	21.8	5.62	0	39	0.7	3.08
	2	40	8,042	22.11	5.97	2	40	0.76	2.92
	3	40	5,399	22.05	5.95	1	38	0.75	2.98
	4	40	5,476	22.33	5.71	2	39	0.72	3.02
	5	40	6,803	22.55	5.61	5	39	0.71	3.02
QR	1	32	8,424	18.84	5.65	1	32	0.8	2.53
	2	32	8,042	18.57	6.19	0	32	0.83	2.55
	3	32	5,399	18.33	6.05	1	32	0.83	2.49
	4	32	5,476	18.59	5.51	2	32	0.78	2.58
	5	32	6,803	18.06	5.27	1	32	0.75	2.63
AR	1	50	8,424	32.02	7.87	4	50	0.83	3.24
	2	50	8,042	32.43	8.31	4	50	0.85	3.22
	3	50	5,399	31.96	8.27	1	50	0.85	3.2
	4	50	5,476	31.69	7.9	5	50	0.83	3.26
	5	50	6,803	32.26	7.66	3	50	0.82	3.25
DM	1	26	8,424	19.18	5.72	3	34	0.76	2.8
	2	26	8,042	17.73	5.58	1	34	0.73	2.9
	3	26	5,399	19.09	6.14	1	33	0.79	2.81
	4	26	5,476	19.19	6.03	1	33	0.79	2.76
	5	26	6,803	18.41	5.61	2	33	0.76	2.75

Candidates receive a scaled score for each cognitive subtest; therefore, scaled score reliabilities and *SEM*s are also provided in Table 19. Unlike the raw score reliability – in which the reliability index (Cronbach's alpha) was generated based on the intercorrelations or internal consistency among the items – the overall reliability of the scaled scores depends on the conditional reliability at each scaled score point instead of on item scores. For this reason, the two reliability indices (Cronbach's alpha and marginal reliability of scaled scores) are not directly comparable.

Table 19 also contains the ranges and means of reliabilities and *SEM*s for the total scaled score. These values were computed as a composite function of the *SEM*s and reliabilities of the cognitive test forms contributing to the total. That is, each total scaled score is a simple sum of the forms of the cognitive tests that were administered to a given candidate. There were five combinations of cognitive test forms and therefore there were five estimates of total scaled score reliability and *SEM*. Reliability for the five forms ranged from 0.89 to 0.92; therefore, the mean reliability for total scaled score was 0.90, reflecting good overall reliability. The mean *SEM* was 83.2, which is very reasonable for the range of total scaled score.

Table 19. Scaled Score Reliability and *SEM* for Cognitive Subtests

Subtest	Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Scaled Score Reliability	*SEM*
VR	1	40	8,424	565	71	300	890	0.7	39.13
	2	40	8,042	570	78	300	900	0.76	38.30
	3	40	5,399	569	76	300	870	0.74	38.87
	4	40	5,476	572	73	300	890	0.72	38.79
	5	40	6,803	574	72	320	890	0.7	39.28
QR	1	32	8,424	667	76	330	900	0.78	35.75
	2	32	8,042	667	87	300	900	0.81	37.72
	3	32	5,399	664	83	330	900	0.8	37.22
	4	32	5,476	665	74	390	900	0.76	36.16
	5	32	6,803	656	68	330	900	0.74	34.92
AR	1	50	8,424	651	90	300	900	0.81	39.36
	2	50	8,042	657	99	300	900	0.84	39.43
	3	50	5,399	652	97	300	900	0.84	38.62
	4	50	5,476	647	90	300	900	0.82	38.18
	5	50	6,803	655	89	300	900	0.81	38.87
DM	1	26	8,424	632	89	330	900	0.73	46.14
	2	26	8,042	610	83	300	900	0.66	48.39
	3	26	5,399	631	93	300	890	0.75	46.57
	4	26	5,476	632	91	300	890	0.72	48.41
	5	26	6,803	623	87	300	890	0.7	47.85
Total	1	148	8,424	2,515	263	1,460	3,460	0.9	83.12
	2	148	8,042	2,504	278	1,410	3,440	0.91	83.26
	3	148	5,399	2,515	289	1,420	3,490	0.92	81.82
	4	148	5,476	2,517	267	1,550	3,410	0.9	84.4
	5	148	6,803	2,508	251	1,440	3,400	0.89	83.38

Item Analysis – Cognitive Subtests

Since 2007, the item development and pretesting plan has been implemented in order to strengthen the UCAT item pool. Improvement of the active item pool is achieved through rounds of item writing, pretesting, data analysis and statistical screening. Each year, new items are developed through item-writing workshops. These newly developed items are then pretested with operational items. At the end of each testing window, both operational and pretest items are analysed. The purpose of item analysis is to examine the item quality and determine whether items are suitable for future use.

Test Analysis – SJT

The raw score means, standard deviations, ranges, internal consistency reliabilities and *SEM*s for each form of the SJT are summarised in Table 20. The test statistics are computed based on all candidates who took the SJT. The maximum number of available score points is 242; however, it has varied in previous years. Therefore, the mean raw score as a percentage of the maximum available score is used to compare the raw score. The mean percentage raw score ranges from 72% to 73%. The reasonably high percent-correct and skewed scaled score distribution indicates that the SJT is capable of identifying the weakest candidates.

Raw score reliabilities for the five SJT forms ranged from 0.79 to 0.81. The reliabilities for all SJT forms are good and comparable to 2019. The *SEM* was based on the raw score metric and ranged from 8.93 to 9.42.

Table 20. SJT Raw Score Test Statistics (all candidates)

Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Mean Percent Raw Score	Alpha	*SEM*
1	63	8,424	173.61	20.59	26	219	0.72	0.79	9.33
2	63	8,042	174.22	20.60	33	220	0.72	0.79	9.42
3	63	5,399	176.32	20.50	16	219	0.73	0.79	9.42
4	63	5,476	175.30	20.63	40	222	0.72	0.81	8.93
5	63	6,803	176.67	19.82	16	221	0.73	0.79	9.08

The band that candidates receive for the SJT is based on their SJT scaled score. Test statistics for scaled scores are provided in Table 21. The scaled scores are linearly related to the raw scores and therefore the raw score reliability applies equally to the scaled scores. This contrasts with the cognitive tests, where the scaled scores are a transformation of the IRT ability values.

Table 21. SJT Scaled Score Test Statistics (all candidates)

Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	*SEM*
1	63	8,424	610.72	70.94	300	768	32.51
2	63	8,042	610.86	68.38	300	764	31.33
3	63	5,399	618.7	71.14	300	769	32.6
4	63	5,476	614.89	70.74	300	776	30.83
5	63	6,803	613.83	67.4	300	766	30.89

Item Analysis – SJT

Each year, new SJT items are developed and reviewed. The SJT items are analysed using classical test theory. A review of the SJT following the 2013 test window showed that an IRT approach is not appropriate given the current polytomous scoring approach used for the SJT. Unlike IRT, classical test statistics are sample-dependent, meaning that they are calculated based on the sample of candidates who respond to each item and are not linked back to a common benchmark group.

Differential Item Functioning

Introduction

Differential Item Functioning (DIF) refers to the potential for items to behave differently for different groups. DIF is generally an undesirable characteristic of an examination because it means that the test is measuring not only the construct it was designed to measure but also an additional characteristic or characteristics of performance that depend on classification or membership in a group, usually a gender or ethnic group classification.

This section describes the methods used to detect DIF for the UCAT examination and provides the results for the 2020 administration.

Detection of DIF

The Mantel–Haenszel procedure was used to detect DIF in the cognitive subtests. It compares reference and focal group performance for candidates within the same ability strata. If there are overall differences between reference group and focal group performance for candidates of the same ability levels, then the item may not be fitting the psychometric model and may be measuring something other than what it was designed to measure.

A hierarchical regression approach using the equated scaled score was used to detect DIF for the SJT. For each comparison, the first column indicates the size of increase in the variance in item responses explained by the regression equation when the group membership variable and an interaction variable of group membership with SJT score were added to the equation. No value is provided where the change did not reach statistical significance at the 95% level.

Criteria for Flagging Items

DIF items were classified into one of three categories: A, B, or C. Category A contains items with negligible DIF, Category B contains items with slight to moderate DIF, and Category C contains items with moderate to large DIF. For the cognitive subtests, these categories are derived from the DIF classification categories developed by Educational Testing Service (ETS) and are defined below:

A: MH D-DIF is not significantly different from zero or has an absolute value <1.0

B: MH D-DIF is significantly different from zero and has an absolute value >=1.0 and <1.5

C: MH-D-DIF is significantly larger than 1.0 and has an absolute value >=1.5.

Items flagged in Category C are typically subjected to further scrutiny. Items flagged in Categories A and B are not reviewed because of the minor statistical significance. The principal interpretation of Category C items is that, based on the present samples, items flagged in this category appear to be functioning differently for the reference and focal groups under comparison. If an item functions differently for two different groups, then content experts may (or may not) be able to determine from the item itself whether the item text contains language or content that may create a bias for the reference or focal group. Therefore, Category C flagging for DIF is necessary but not sufficient grounds for revision and possible removal of the item from the pools.

For the SJT, effects that explain less than 1% of score variance (R squared change <0.01) are considered negligible for flagging purposes, and items that do not reach significance or explain less than this proportion of variance are labelled 'A', meaning that they can be considered free of DIF. Larger effects, where the group variable has a significant beta coefficient, are labelled 'B' or 'C'. Changes of 0.01 or above are considered slight to moderate and labelled 'B', unless all of the change is explained by the interaction term, in which case they are labelled 'A'. Changes above 0.05 (5% of the variance in responses) are considered moderate to large and are labelled 'C' where there is a significant main effect of the group difference variable.

All items with moderate to large DIF were reviewed and dropped from the operational item pool where any potential unfairness in the content was identified.

Comparison Groups for DIF Analysis

DIF analyses were conducted for the pretest and operational items when sample sizes were large enough. The UCAT DIF comparison groups are based on gender, age, ethnicity, socio-economic status, level of education, first language, and permanent residence. Age was separated into groups of less than 20 years old and greater than 35 years old. There are 17 ethnic categories in the UCAT database. For the DIF analyses, several of these categories were collapsed into meaningful, broader groups. The DIF ethnic categories used for these analyses (collapsed where indicated) were as follows:

White:	White – British
Black:	Black – Black/British – African, Black – Black/British – Caribbean, Black – Black/British Other
Asian:	Asian – Asian/British – Bangladeshi, Asian – Asian/British – Indian, Asian – Asian/British – Other Asian, Asian – Asian/British – Pakistani.
Chinese:	Asian – Asian/British – Chinese
Mixed:	Mixed – Mixed – Other, Mixed – White/Asian, Mixed – White/Black African, Mixed – White/Black Caribbean
Other:	Other ethnic group

Sample Size Requirements

Minimum sample-size requirements used for the UCAT DIF analyses were at least 50 focal group candidate responses and at least 200 total (focal plus reference) candidate responses. Because pretest items were distributed across multiple versions of the forms, fewer responses were available per item than for operational items. As a result, it was not possible to compute DIF for many of the pretest items for some group comparisons.

DIF Results – Cognitive Subtests

Table 22 (operational items) and Table 23 (pretest items) in Appendix A show the number and percentages of items classified into each of the three DIF categories along with the quantities for which insufficient data were available to compute DIF (Category NA).

In operational DIF analysis, comparisons between all variables except age groups met sample size requirements to compute DIF. For age groups, 39% to 41% of items did not meet sample size requirements. For pretest items, comparisons between age groups, NS-SEC categories and ethnic groups failed to meet the minimum sample size requirements. These items will be re-evaluated for DIF when they are used in future operational forms.

For the operational pools, there were 17 occurrences of Category C DIF across all cognitive subtests and comparisons. The proportion of Category C DIF out of all possible comparisons across the four cognitive tests was extremely low. Of these 17 occurrences, six related to age group, two to level of education, eight to ethnicity, and one to gender. For the pretest items, there were three occurrences of Category C DIF, one in each of the First Language, Residence and Education Level comparison groups. It should be noted that as pretest items are seen by fewer candidates, a significant number of comparisons could not be made due to low sample numbers in the focal groups. Taken together, the results indicated very little DIF occurrence in the UCAT items.

DIF Results – SJT

Table 24 (operational items) and Table 25 (pretest items) in Appendix A show the number and percentages of items classified into each of the three DIF categories along with the quantities for which insufficient data were available to compute DIF (category NA<200).

In operational DIF analysis, all items met sample size requirements to compute DIF for all comparison groups for the SJT. For some pretest items, comparisons between White and Black, White and Chinese, and White and Mixed, did not meet minimal sample size requirements. These items will be re-evaluated for DIF when they are used in future operational forms.

For the operational SJT pool, there were no occurrences of Category C DIF, and 39 instances of Category B DIF overall. For the pretest items, there were five occurrences of Category C DIF. It should be noted that as pretest items are seen by fewer candidates, a significant number of comparisons could not be made due to low sample numbers in the focal groups.

Appendix A. Cognitive Subtest DIF Summary Tables

Table 22. DIF Classification: Operational Pool

Comparison Group	MH D-DIF Code	VR		QR		AR		DM	
		N Items	%	*N* Items	%	*N* Items	%	*N* Items	%
Male/Female	A	199	100%	157	100%	250	100%	127	98%
	B	1	0%	0	0%	0	0%	2	2%
	C	0	0%	0	0%	0	0%	1	1%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
Age <20/>35	A	119	60%	96	61%	149	60%	72	55%
	B	0	0%	0	0%	0	0	0	0%
	C	0	0%	0	0%	1	0	5	4%
	NA	81	40%	61	39%	100	40%	53	41%
	Total	200	100%	157	100%	250	100%	130	100%
White/Black	A	197	98%	154	98%	247	99%	118	91%
	B	3	2%	3	2%	2	1%	8	6%
	C	0	0%	0	0%	1	0%	4	3%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
White/Asian	A	196	98%	155	99%	249	100%	126	97%
	B	4	2%	1	1%	1	0%	4	3%
	C	0	0%	1	1%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
White/ Chinese	A	199	100%	157	100%	250	100%	129	99%
	B	0	0%	0	0%	0	0%	0	0%
	C	1	0%	0	0%	0	0%	1	1%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
White/Mixed	A	200	100%	157	100%	250	100%	130	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
NS-SEC Class 1/2	A	200	100%	157	100%	250	100%	130	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
NS-SEC Class 1/3	A	200	100%	157	100%	250	100%	130	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
NS-SEC Class 1/4	A	200	100%	157	100%	250	100%	130	100%
	B	0	0%	0	0%	0	0%	0	0%

Comparison Group	MH D-DIF Code	VR		QR		AR		DM	
		N Items	%	*N* Items	%	*N* Items	%	*N* Items	%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
NS-SEC Class 1/5	A	199	100%	157	100%	250	100%	130	100%
	B	1	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
Below Honours Degree/ Honours Degree	A	199	100%	157	100%	250	100%	121	93%
	B	1	0%	0	0%	0	0%	7	5%
	C	0	0%	0	0%	0	0%	2	2%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
English/ Non-English 1st Language	A	200	100%	157	100%	250	100%	130	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%
UK/ Non-UK	A	200	100%	155	99%	250	100%	129	99%
	B	0	0%	2	1%	0	0%	1	1%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	157	100%	250	100%	130	100%

Note. NA: Insufficient data to compute MH D-DIF

Table 23. DIF Classification: Pretest Pool

Comparison Group	MH D-DIF Code	VR		QR		AR		DM	
		N Items	%	*N* Items	%	*N* Items	%	*N* Items	%
Male/Female	A	224	100%	217	100%	290	100%	247	100%
	B	0	0%	1	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	224	100%	218	100%	290	100%	247	100%
Age <20/>35	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	224	100%	218	100%	290	100%	247	100%
	Total	224	100%	218	100%	290	100%	247	100%
White/Black	A	141	63%	139	64%	144	50%	21	9%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	83	37%	79	36%	146	50%	226	91%
	Total	224	100%	218	100%	290	100%	247	100%

Comparison Group	MH D-DIF Code	VR		QR		AR		DM	
		N Items	%	*N* Items	%	*N* Items	%	*N* Items	%
White/Asian	A	224	100%	218	100%	290	100%	206	83%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	41	17%
	Total	224	100%	218	100%	290	100%	247	100%
White/ Chinese	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	224	100%	218	100%	290	100%	247	100%
	Total	224	100%	218	100%	290	100%	247	100%
White/Mixed	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	224	100%	218	100%	290	100%	247	100%
	Total	224	100%	218	100%	290	100%	247	100%
NS-SEC Class 1/2	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	224	100%	218	100%	290	100%	247	100%
	Total	224	100%	218	100%	290	100%	247	100%
NS-SEC Class 1/3	A	1	0%	0	0%	2	1%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	223	100%	218	100%	288	99%	247	100%
	Total	224	100%	218	100%	290	100%	247	100%
NS-SEC Class 1/4	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	224	100%	218	100%	290	100%	247	100%
	Total	224	100%	218	100%	290	100%	247	100%
NS-SEC Class 1/5	A	0	0%	2	1%	5	2%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	224	100%	216	99%	285	98%	247	100%
	Total	224	100%	218	100%	290	100%	247	100%
Below Honours Degree/Honours Degree	A	224	100%	217	100%	290	100%	247	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	1	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	224	100%	218	100%	290	100%	247	100%
English/Non-English 1st Language	A	224	100%	218	100%	289	100%	247	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	1	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	224	100%	218	100%	290	100%	247	100%
	A	223	100%	218	100%	290	100%	224	91%

Comparison Group	MH D-DIF Code	VR		QR		AR		DM	
		N Items	%	*N* Items	%	*N* Items	%	*N* Items	%
UK/ Non-UK	B	0	0%	0	0%	0	0%	1	0%
	C	1	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	22	9%
	Total	224	100%	218	100%	290	100%	247	100%

Note. NA: Insufficient data to compute MH D-DIF

Table 24. SJT DIF Classification: Operational Pool

Comparison Group	Degree of DIF					
	A		B		C	
	N Items	%	*N* Items	%	*N* Items	%
Male/Female	196	99%	2	1%	0	0%
Age <20/>35	195	98%	3	2%	0	0%
White/Black	187	94%	11	6%	0	0%
White/Asian	186	94%	12	6%	0	0%
White/Chinese	196	99%	2	1%	0	0%
White/Mixed	198	100%	0	0%	0	0%
NS-SEC Class 1/2	198	100%	0	0%	0	0%
NS-SEC Class 1/3	198	100%	0	0%	0	0%
NS-SEC Class 1/4	198	100%	0	0%	0	0%
NS-SEC Class 1/5	198	100%	0	0%	0	0%
UK/Non-UK	193	97%	5	3%	0	0%
English 1st Language/Other 1st Language	194	98%	4	2%	0	0%
Graduate/Non-Graduate	198	100%	0	0%	0	0%
Delivery Method	198	100%	0	0%	0	0%
Medicine/Dentistry	198	100%	0	0%	0	0%

Table 25. SJT DIF Classification: Pretest Pool

Comparison Group	Degree of DIF							
	A		B		C		N<200	
	N Items	%	*N* Items	%	*N* Items	%	*N* Items	%
Male/Female	229	95.42%	11	4.58%	0	0.00%	0	0.00%
Age <20/>35	230	95.83%	9	3.75%	1	0.42%	0	0.00%
White/Black	158	65.83%	9	3.75%	3	1.25%	70	29.17%
White/Asian	224	93.33%	16	6.67%	0	0.00%	0	0.00%
White/Chinese	96	40.00%	2	0.83%	0	0.00%	142	59.17%
White/Mixed	122	50.83%	3	1.25%	0	0.00%	115	47.92%
NS-SEC Class 1/2	240	100.00%	0	0.00%	0	0.00%	0	0.00%
NS-SEC Class 1/3	236	98.33%	3	1.25%	1	0.42%	0	0.00%
NS-SEC Class 1/4	238	99.17%	2	0.83%	0	0.00%	0	0.00%

NS-SEC Class 1/5	234	97.50%	6	2.50%	0	0.00%	0	0.00%
UK/Non-UK	231	96.25%	9	3.75%	0	0.00%	0	0.00%
English 1st Language/Other 1st Language	224	93.33%	16	6.67%	0	0.00%	0	0.00%
Graduate/Non-Graduate	230	95.83%	10	4.17%	0	0.00%	0	0.00%
Delivery Method	235	97.92%	5	2.08%	0	0.00%	0	0.00%
Medicine/Dentistry	236	98.33%	4	1.67%	0	0.00%	0	0.00%