University Clinical Aptitude Test (UCAT) Consortium
UCAT Examination

Executive Summary
Testing Interval: 1 July 2019 – 2 October 2019

Prepared by:
Pearson VUE
June 2020

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.

Contents

List of Tables

Executive Summary

The University Clinical Aptitude Test (UCAT) was administered in 2019 from 1 July to 2 October. During this period, a total of 29,366 exams were administered. Each exam consisted of four 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).

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 6 were pretest. The exam was administered via computer in a 120-minute time period including administration of instructions. Each of the five sections was timed separately. There were four groups of candidates who received time accommodation in 2019. 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). Results were provided to the candidates at the conclusion of testing and then later sent to the schools to which the candidates had applied.

Candidate performance was broadly consistent with previous years. The average VR and DM scores were slightly lower by 2 points and 6 points respectively. The average QR and AR scores were slightly higher by 4 points and 1 point respectively. The SJT band distribution was broadly similar to that observed in 2018.

In terms of candidate performance by social-economic group (UK candidates only; based on parental profession), Category 1 (Managerial and Professional Occupations) was consistently associated with higher mean scaled scores in the cognitive subtests. The lowest average subtest scaled scores occurred for Category 5 (Semi-routine or Routine Occupations) for all subtests. These social-economic trends are similar to those in 2018.

Candidate age was broken into five groups: ≤15, 16 to 19, 20 to 24, 25 to 34, and ≥35. Performance across various age groups was examined separately by the candidates' highest educational qualification. For candidates with Honours degrees, the age group 20 to 24 showed the highest scores across all cognitive sections, which is consistent with the results in 2018. For candidates with school-leaving qualifications (i.e., below Honours degrees), the age group 16 to 19 had the highest scores, which also reflects the 2018 results.

The report also includes the performance analysis by the candidates' first language (English vs. non-English for UK and non-UK residents). The results indicated 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. This is consistent with the 2018 cohort. The SJT showed similar trends to the cognitive sections by first language.

Background

The University Clinical Aptitude Test (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 good clinicians. 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 2019 testing period began on 1 July and ended on 2 October. During this period, a total of 29,366 exams were administered. Five forms each of the Verbal Reasoning (VR), Quantitative Reasoning (QR), Abstract Reasoning (AR), Decision Making (DM), and Situational Judgement Test (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 cognitive tests and a set of pretest (unscored) items.

Table 1. Composition of the Five UCAT Forms

UCAT Form	Verbal Reasoning	Quantitative Reasoning	Abstract Reasoning	Decision Making	Situational Judgement
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 2018 administrations and also from items that had been pretested during these years. The SJT items were developed from operational and pretest items used from 2013 to 2018. All items (operational and pretest) were analysed, and those with acceptable item statistics were saved as the active item bank.

Design of Exam

The UCAT is an aptitude exam and is designed to measure innate cognitive abilities in addition to individuals' judgements 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 2019 exam contained one SJT subtest and four scored cognitive subtests: VR, QR, AR and DM. All sections contained both operational (scored) and pretest (unscored) items. 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 2019. 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 2 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

Students' 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 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. There were 29,366 candidate scores collected during the 2019 testing window.

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

Test	Total *N*	Mean	*SD*	Min	Max
VR	29,366	565.2	75.57	300	900
QR	29,366	662.17	76.45	330	900
AR	29,366	638.04	85.95	300	900
DM	29,366	617.7	76.85	300	890
Total	29,366	2,483.11	248.56	1,400	3,510

Table 4 summarises the scaled score statistics for UCAT non-SEN candidates and SEN candidates. SEN candidates were allocated additional time and 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	27,993	564.29	75.49	300	900
	QR	27,993	661.14	76.16	330	900
	AR	27,993	636.47	85.19	300	900
	DM	27,993	617.27	76.76	300	890
	Total	27,993	2,479.18	247.75	1,400	3,510
UCATSEN	VR	1,162	583.07	74.22	370	890
	QR	1,162	682.95	78.95	460	900
	AR	1,162	669.77	94.47	300	900
	DM	1,162	625.6	77.68	300	890
	Total	1,162	2,561.39	248.61	1,520	3,380
UCATSENSA	VR	103	596.41	79.11	450	830
	QR	103	680.1	85.43	460	900
	AR	103	677.96	86.12	490	870
	DM	103	633.11	81	380	770
	Total	103	2,587.57	259.31	1,960	3,220
UCATSEN50	VR	61	584.75	78.41	320	890
	QR	61	695.41	76.35	500	860

Exam	Test	Total *N*	Mean	*SD*	Min	Max
	AR	61	670	115.2	300	870
	DM	61	623.93	87.56	380	890
	Total	61	2,574.1	290.83	1,500	3,440
UCATSA	VR	47	573.4	73.43	450	790
	QR	47	676.6	78.36	540	880
	AR	47	657.23	90.74	400	860
	DM	47	634.47	76.61	480	850
	Total	47	2,541.7	256.12	1,960	3,170

Situational Judgement Test

For the Situational Judgement Test candidates are awarded one of four bands to reflect their performance on the operational items in the SJT. 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	662-900	22%	Those in Band 1 demonstrated an excellent level of performance, showing similar judgement in most cases to the panel of experts.
Band 2	597-661	38%	Those in Band 2 demonstrated a good, solid level of performance, showing appropriate judgement frequently, with many responses matching model answers.
Band 3	512-596	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	300-511	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 29,366 candidates who took the UCAT during the 2019 testing window. The proportions observed in the 2019 SJT are similar to the intended percentages. Band 1 is somewhat lower than intended and Bands 2-4 are slightly higher.

Table 6. SJT Band Distribution in 2019

SJT Band	Number of Candidates	Percentage of Candidates
Band 1	4,985	17%
Band 2	11,640	39.6%
Band 3	9,658	32.9%
Band 4	3,083	10.5%
Total	29,366	100%

Table 7 summarises the percentage by band and the scaled score statistics for SEN and non-SEN candidates. UCATSEN candidates outperformed non-SEN candidates on the SJT. Low candidate volumes for the other SEN exams prevent conclusions from being drawn from their band distributions.

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	27,993	16.8%	39.4%	33%	10.8%	597.66	69.42	300	770
UCATSEN	1,162	22.3%	43.2%	29.6%	4.9%	614.68	60.64	300	766
UCATSENSA	103	16.5%	49.5%	30.1%	3.9%	610.15	53.44	474	735
UCATSEN50	61	16.4%	44.3%	31.1%	8.2%	602.62	72.33	300	738
UCATSA	47	17%	51.1%	29.8%	2.1%	623.34	58.97	376	745

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. However, these scores are not issued to candidates and are not directly comparable to the cognitive subtests scaled scores.

Gender

Table 8 presents scaled score summary statistics for males and females for each of the subtests. Females constituted 18,646 (64%) candidates and males 10,628 (36%). On average, males slightly outperformed females on VR, QR, AR and DM. Female candidates outperformed male candidates on the SJT, as observed in previous years.

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

Test	Gender	Total *N*	Total %	Mean	*SD*	Min	Max
Verbal Reasoning[1]	Female	18,646	64%	561.91	74.70	300	900
	Male	10,628	36%	570.95	76.67	300	900
Quantitative Reasoning[2]	Female	18,646	64%	653.33	73.76	330	900
	Male	10,628	36%	677.65	78.58	390	900
Abstract Reasoning[3]	Female	18,646	64%	635.41	84.88	300	900
	Male	10,628	36%	642.65	87.63	300	900
Decision Making[4]	Female	18,646	64%	614.68	76.77	300	890
	Male	10,628	36%	623.00	76.70	300	890
Total Cognitive Scaled Score[5]	Female	18,646	64%	2,465.33	245.61	1,400	3,440
	Male	10,628	36%	2,514.26	250.62	1,410	3,510
Situational Judgement Test[6]	Female	18,646	64%	603.08	67.40	300	766
	Male	10,628	36%	590.21	71.34	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 = 9.87 (df= 29272, p<0.01) assuming equal variance.
[2] T-statistics = 26.49 (df= 29272, p<0.01) assuming equal variance.
[3] T-statistics = 6.94 (df= 29272, p<0.01) assuming equal variance.
[4] T-statistics = 8.92 (df= 29272, p<0.01) assuming equal variance.
[5] T-statistics = 16.27 (df= 29272, p<0.01) assuming equal variance.
[6] T-statistics = -15.51 (df= 29272, 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
Verbal Reasoning	Non-UK	5,729	20%	545.28	77.31	300	890
	UK - Asian	8,759	30%	554.75	68.45	300	890
	UK - Black	2,434	8%	541.56	66.92	320	830
	UK - Chinese	375	1%	576.72	77.67	380	870
	UK - Mixed Race	1,299	4%	578.28	76.20	300	900
	UK - Other	1,264	4%	542.77	74.39	300	830
	UK - White	9,506	32%	593.64	73.67	300	900
Quantitative Reasoning	Non-UK	5,729	20%	648.81	82.37	330	900
	UK - Asian	8,759	30%	662.99	75.08	330	900
	UK - Black	2,434	8%	626.76	66.02	390	880
	UK - Chinese	375	1%	705.33	83.39	500	900
	UK - Mixed Race	1,299	4%	667.19	73.21	480	900
	UK - Other	1,264	4%	644.94	72.74	430	900
	UK - White	9,506	32%	678.43	71.67	390	900
Abstract Reasoning	Non-UK	5,729	20%	620.01	90.05	300	890
	UK - Asian	8,759	30%	643.07	85.14	300	900
	UK - Black	2,434	8%	607.06	81.05	300	900
	UK - Chinese	375	1%	671.28	87.03	380	900
	UK - Mixed Race	1,299	4%	646.86	85.59	300	890
	UK - Other	1,264	4%	629.56	82.21	300	900
	UK - White	9,506	32%	650.81	81.76	300	900
Decision Making	Non-UK	5,729	20%	603.90	82.17	300	890
	UK - Asian	8,759	30%	608.12	73.07	300	890
	UK - Black	2,434	8%	585.93	73.29	300	840
	UK - Chinese	375	1%	637.09	71.81	440	890
	UK - Mixed Race	1,299	4%	629.64	74.79	310	840
	UK - Other	1,264	4%	595.06	77.34	300	880
	UK - White	9,506	32%	643.60	69.94	310	890
Total Cognitive Scaled Score	Non-UK	5,729	20%	2,418.00	265.42	1,400	3,440
	UK - Asian	8,759	30%	2,468.93	237.34	1,520	3,300
	UK - Black	2,434	8%	2,361.31	222.44	1,480	3,230
	UK - Chinese	375	1%	2,590.43	251.34	1,900	3,510
	UK - Mixed Race	1,299	4%	2,521.96	240.85	1,690	3,350
	UK - Other	1,264	4%	2,412.33	247.22	1,480	3,110
	UK - White	9,506	32%	2,566.48	224.74	1,460	3,380
	Non UK	5729	20%	569.35	80.03	300	759
	UK - Asian	8759	30%	594.75	65.67	300	749

Test	Ethnic Group	Total *N*	Total %	Mean	*SD*	Min	Max
Situational Judgement Test	UK - Black	2434	8%	588.97	68.18	300	751
	UK - Chinese	375	1%	607.81	62.13	433	743
	UK - Mixed Race	1299	4%	611.29	63.23	305	749
	UK - Other	1264	4%	591.70	73.74	300	749
	UK - White	9506	32%	620.53	57.03	300	770

The UK-White ethnic category made up 32% of the testing population. Proportions for the other ethnic groups ranged from 1% to 30%. There was considerable variation in means among the different ethnic groups. For VR and DM, the highest-performing group was UK-White. For QR and AR, the highest-performing group was UK-Chinese. This is consistent with the 2018 exam. UK-Black performed worst on all cognitive subtests.

NS-SEC

Table 10 provides scaled score summary statistics for all UK candidates by NS-SEC class (occupation and employment status). 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 2 had the highest mean and Class 5 had the lowest.

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

Test	NS-SEC Group	Total *N*	Total %	Mean	*SD*	Min	Max
Verbal Reasoning	1	15,049	64%	580.26	74.54	300	900
	2	1,049	4%	575.42	74.21	320	870
	3	1,455	6%	552.91	69.45	360	870
	4	719	3%	552.45	70.93	320	790
	5	1,521	6%	544.14	63.51	340	830
	NA	3844	16%	548.51	71.17	300	890
Quantitative Reasoning	1	15,049	64%	674.40	74.45	390	900
	2	1,049	4%	663.18	72.05	430	900
	3	1,455	6%	652.39	71.62	390	900
	4	719	3%	648.92	68.26	330	880
	5	1,521	6%	646.06	70.33	430	880
	NA	3844	16%	646.46	73.37	390	900
Abstract Reasoning	1	15,049	64%	650.74	84.41	300	900
	2	1,049	4%	635.69	80.59	300	880
	3	1,455	6%	630.30	80.80	320	890
	4	719	3%	626.12	79.12	380	900
	5	1,521	6%	625.04	80.06	300	900
	NA	3844	16%	626.13	84.45	300	890
Decision Making	1	15,049	64%	631.82	73.75	300	890
	2	1,049	4%	624.44	73.02	310	880
	3	1,455	6%	604.10	69.70	310	850
	4	719	3%	601.81	77.84	380	850
	5	1,521	6%	595.17	68.50	300	880
	NA	3844	16%	598.22	75.63	300	840

Test	NS-SEC Group	Total *N*	Total %	Mean	*SD*	Min	Max
Total Cognitive Scaled Score	1	15,049	64%	2,537.22	237.72	1,480	3,510
	2	1,049	4%	2,498.74	233.56	1,680	3,350
	3	1,455	6%	2,439.71	222.30	1,700	3,250
	4	719	3%	2,429.29	233.22	1,520	3,300
	5	1,521	6%	2,410.41	214.95	1,790	3,110
	NA	3844	16%	2419.32	241.95	1460	3270
Situational Judgement Test	1	15,049	64%	611.54	61.22	300	770
	2	1,049	4%	612.13	62.21	306	749
	3	1,455	6%	594.90	65.02	322	735
	4	719	3%	593.27	69.02	300	742
	5	1,521	6%	592.50	64.22	300	746
	NA	3844	16%	591.35	71.08	300	751

Note. Codes for NS-SEC Groups
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 level and Honours Degree level or above. Candidates in the Honours Degree level or above category were mostly in the 20 to 24 age group, which also represented the highest mean scores across all four cognitive subtests. Candidates in the Below Honours Degree level category were mostly in the 16 to 19 age group, which showed the highest mean scores across all four cognitive subtests.

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
Verbal Reasoning	Below Honours degree level	Up to 15	46	0%	543.70	87.01	340	740
		16-19	21,304	96%	566.48	74.27	300	900
		20-24	702	3%	543.18	86.77	300	890
		25-34	170	1%	545.12	80.73	380	870
		>=35	32	0%	502.81	68.97	400	720
	Honours degree level or above	Up to 15	1	0%	NA[a]	NA	NA	NA
		16-19	927	14%	543.75	70.38	320	790
		20-24	4,384	65%	572.56	74.59	300	900
		25-34	1,261	19%	569.45	83.28	300	890
		>=35	196	3%	530.97	87.73	320	790
Quantitative Reasoning		Up to 15	46	0%	622.17	79.64	480	880

Test	Highest Qualification	Age Group	Total *N*	% Total *N*	Mean	*SD*	Min	Max
	Below Honours degree level	16-19	21,304	96%	668.00	75.88	330	900
		20-24	702	3%	636.75	87.10	330	900
		25-34	170	1%	614.41	75.12	430	870
		>=35	32	0%	575.63	60.74	430	680
	Honours degree level or above	Up to 15[a]	1	0%	NA	NA	NA	NA
		16-19	927	14%	648.88	74.18	390	900
		20-24	4,384	65%	656.69	72.23	390	900
		25-34	1,261	19%	638.57	71.96	430	900
		>=35	196	3%	593.57	81.68	390	830
Abstract Reasoning	Below Honours degree level	Up to 15	46	0%	591.30	72.04	380	750
		16-19	21,304	96%	642.14	85.13	300	900
		20-24	702	3%	620.43	96.29	300	890
		25-34	170	1%	592.88	84.13	300	830
		>=35	32	0%	574.06	69.32	460	710
	Honours degree level or above	Up to 15	1	0%	NA	NA	NA	NA
		16-19	927	14%	631.95	86.91	300	890
		20-24	4,384	65%	638.19	83.94	300	900
		25-34	1,261	19%	613.95	82.52	300	880
		>=35	196	3%	560.41	95.12	300	890
Decision Making	Below Honours degree level	Up to 15	46	0%	600.00	72.66	390	770
		16-19	21,304	96%	622.85	75.35	300	890
		20-24	702	3%	585.48	88.16	300	890
		25-34	170	1%	566.12	87.13	300	790
		>=35	32	0%	540.63	68.11	380	650
	Honours degree level or above	Up to 15	1	0%	NA	NA	NA	NA
		16-19	927	14%	603.54	76.42	380	880
		20-24	4,384	65%	615.71	73.72	310	890
		25-34	1,261	19%	597.96	79.39	300	840
		>=35	196	3%	556.12	91.06	310	800
Total Score	Below Honours degree level	Up to 15	46	0%	2,357.17	250.96	1,760	3,020
		16-19	21,304	96%	2,499.47	243.81	1,440	3,510
		20-24	702	3%	2,385.84	297.39	1,400	3,440
		25-34	170	1%	2,318.53	266.71	1,520	3,030
		>=35	32	0%	2,193.13	214.53	1,870	2,710
	Honours degree level or above	Up to 15	1	0%	NA	NA	NA	NA
		16-19	927	14%	2,428.12	245.04	1,600	3,250
		20-24	4,384	65%	2,483.15	235.96	1,430	3,380
		25-34	1,261	19%	2,419.94	251.72	1,410	3,320
		>=35	196	3%	2,241.07	300.38	1,460	3,140
Situational Judgement	Below Honours degree level	Up to 15	46	0%	544.52	76.69	301	668
		16-19	21,304	96%	595.51	66.86	300	770
		20-24	702	3%	581.08	86.66	300	739
		25-34	170	1%	588.08	84.53	328	759

Test	Highest Qualification	Age Group	Total *N*	% Total *N*	Mean	*SD*	Min	Max
		>=35	32	0%	581.63	74.32	405	698
	Honours degree level or above	Up to 15	1	0%	NA	NA	NA	NA
		16-19	927	14%	574.74	75.10	300	746
		20-24	4,384	65%	621.57	62.07	300	760
		25-34	1,261	19%	616.46	71.34	300	763
		>=35	196	3%	578.21	102.90	300	745

[a]There was only 1 candidate in this category. For confidentiality, these scores are not reported.

Similar to cognitive subtests, the Below Honours Degree level had the highest mean SJT scaled scores at ages 16 to 19 and the Honours Degree level or above category had the highest mean SJT score at ages 20 to 24. These trends are consistent with those observed in 2017 and in 2018.

First Language

Scaled score analysis by the candidates' 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 four cognitive sections compared to candidates whose first language is not English for both UK and non-UK candidates. UK candidates outperformed non-UK candidates.

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
Verbal Reasoning	UK	English	16,762	57%	582.39	73.1	300	900
		Other	6,875	23%	539.91	68.53	300	870
	Non-UK	English	2,418	8%	569.2	80.56	300	890
		Other	3,311	11%	527.81	69.86	300	870
Quantitative Reasoning	UK	English	16,762	57%	673.06	73.24	330	900
		Other	6,875	23%	646.75	74.53	390	900
	Non-UK	English	2,418	8%	660.79	82.69	330	900
		Other	3,311	11%	640.06	81.04	330	900
Abstract Reasoning	UK	English	16,762	57%	647.78	83.58	300	900
		Other	6,875	23%	629.32	84.77	300	900
	Non-UK	English	2,418	8%	624.92	90.17	300	890
		Other	3,311	11%	616.42	89.81	300	890
Decision Making	UK	English	16,762	57%	632.6	72.54	300	890
		Other	6,875	23%	592.88	73.89	300	880
	Non-UK	English	2,418	8%	619.88	81.79	300	890
		Other	3,311	11%	592.23	80.48	300	880
Total Score	UK	English	16,762	57%	2,535.82	233.39	1,460	3,510
		Other	6,875	23%	2,408.85	237.88	1,480	3,250
	Non-UK	English	2,418	8%	2,474.79	267.68	1,410	3,440
		Other	3,311	11%	2,376.52	255.95	1,400	3,380
	UK	English	16,762	57%	613.17	60.32	300	770

Test	Country of Residence	First Language	Total *N*	% Total *N*	Mean	*SD*	Min	Max
Situational Judgement		Other	6,875	23%	586.71	69.5	300	763
	Non-UK	English	2,418	8%	586.5	72.21	300	759
		Other	3,311	11%	556.82	83.1	300	756

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*) 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 estimation of item difficulty, or *b*, parameter.

Test Analysis: Cognitive Subtests

The raw score means, standard deviations, ranges, internal consistency reliabilities (Cronbach's alpha), and standard errors of measurement for each form of each subtest are summarised in Table 13.

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 13. Raw Score Test Statistics

	Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Alpha	*SEM*
Verbal Reasoning	1	40	5,536	21.27	5.78	1	39	0.73	3.00
	2	40	7,000	21.78	5.55	3	39	0.71	2.99
	3	40	5,599	21.67	5.93	2	39	0.75	2.96
	4	40	5,614	22.45	6	3	40	0.76	2.94
	5	40	5,617	21.65	5.77	3	40	0.73	3.00
Quantitative Reasoning	1	32	5,536	18.36	5.61	1	32	0.79	2.57
	2	32	7,000	18.87	5.76	1	32	0.81	2.51
	3	32	5,599	18.64	5.58	1	32	0.79	2.56
	4	32	5,614	18.3	5.51	2	32	0.78	2.58
	5	32	5,617	18.81	5.63	1	32	0.8	2.52
Abstract Reasoning	1	50	5,536	30.27	7.66	2	50	0.81	3.34
	2	50	7,000	32.02	7.9	0	50	0.82	3.35
	3	50	5,599	30.44	7.52	5	49	0.8	3.36
	4	50	5,614	30.7	7.73	0	50	0.82	3.28
	5	50	5,617	30.39	7.14	6	49	0.78	3.35
Decision Making	1	26	5,536	17.74	4.57	2	32	0.62	2.82
	2	26	7,000	17.32	4.86	1	32	0.69	2.71
	3	26	5,599	17.72	5.02	2	32	0.69	2.8
	4	26	5,614	17.96	4.94	2	32	0.68	2.79
	5	26	5,617	18.49	4.71	3	30	0.66	2.75

Candidates receive a scaled score for each cognitive subtest; therefore, scaled score reliabilities and standard errors are also provided in Table 14. 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. *SEM* also provides information about reliability of the scaled scores. Contrary to reliability coefficients (for which larger numbers reflect more reliable scores), larger standard errors indicate poorer reliability. The *SEM* of the scaled scores averaged 39 for VR, 36 for QR, 38 for AR, and 44 for DM.

Table 14 also contains the ranges and means of reliabilities and standard errors for the total scaled score. These values were computed as a composite function of the standard errors and reliabilities of the cognitive test forms contributing to the total. That is, each total scaled score is a simple sum (linear composite) 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 standard error. Reliability for the five forms ranged from 0.89 to 0.90, therefore the average reliability for the total scaled score was 0.90, reflecting good overall reliability. The average standard error was 79.87, which is very reasonable for the range of total scaled score.

Table 14. Scaled Score Reliability and Standard Error of Measurement for Cognitive Subtests

Tests	Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Scaled Score Reliability	*SEM*
VR	1	40	5,536	557.5	74.21	300	890	0.72	39.27
	2	40	7,000	565.3	70.72	300	890	0.7	38.74
	3	40	5,599	566.0	77.71	300	890	0.74	39.62
	4	40	5,614	573.2	78.97	300	900	0.75	39.48
	5	40	5,617	564.0	76.33	300	900	0.73	39.66
QR	1	32	5,536	657.2	75.93	330	900	0.78	35.61
	2	32	7,000	666.4	79.17	330	900	0.78	37.14
	3	32	5,599	662.9	75.69	330	900	0.77	36.30
	4	32	5,614	657.6	73.41	390	900	0.76	35.96
	5	32	5,617	665.7	76.72	330	900	0.78	35.98
AR	1	50	5,536	632.3	85.32	300	900	0.8	38.16
	2	50	7,000	652.0	92.13	300	900	0.81	40.16
	3	50	5,599	632.9	83.32	300	890	0.79	38.18
	4	50	5,614	636.8	86.45	300	900	0.8	38.66
	5	50	5,617	632.5	78.39	300	890	0.77	37.60
DM	1	26	5,536	609.3	71.55	300	890	0.63	43.52
	2	26	7,000	611.8	78.86	300	890	0.68	44.61
	3	26	5,599	615.1	79.06	300	890	0.69	44.02
	4	26	5,614	626.2	78.54	300	890	0.68	44.43
	5	26	5,617	627.4	73.54	310	840	0.65	43.51
Total[a]	1	148	5,536	2456	241	1410	3320	0.89	80.09
	2	148	7,000	2495	253	1490	3510	0.9	79.96
	3	148	5,599	2477	252	1400	3380	0.9	79.63
	4	148	5,614	2494	252	1410	3350	0.9	79.75
	5	148	5,617	2490	241	1630	3370	0.89	79.89

[a]Based on five combinations of cognitive test forms.

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* for each form of the SJT are summarised in Table 15. The test statistics are computed based on all candidates who took the SJT. The maximum number of available score points is 244 for all forms, 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 percent raw score ranges from 69% on Form 5 to 71% on Form 2 and Form 3. 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 0.78 to 0.81. The reliabilities for all SJT forms are good and comparable to 2018. As expected, the increase in the difficulty of the forms has not impacted on the reliability of the SJT. The *SEM* was based on the raw score metric and ranged from 9.41 to 9.74.

Table 15. SJT Raw Score Test Statistics (all candidates)

Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Mean Percent Raw Score	Alpha	*SEM*
1	63	5,536	172.00	22.00	0	217	70%	0.81	9.64
2	63	7,000	172.32	19.85	0	222	71%	0.78	9.41
3	63	5,599	172.53	20.39	68	222	71%	0.78	9.60
4	63	5,614	171.85	21.33	0	219	70%	0.80	9.50
5	63	5,617	167.59	20.90	4	219	69%	0.78	9.74

The band that candidates receive for the SJT is based on their SJT scaled score. Test statistics for scaled scores are provided in Table 16. The scaled scores are linearly related to the raw scores and therefore the raw score reliability applies equally to the scaled scores. This is in contrast to the cognitive tests where the scaled scores are a transformation of the IRT ability values.

Table 16. SJT Scaled Score Test Statistics (all candidates)

Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	*SEM*
1	63	5,536	600.95	72.03	300	749	31.55
2	63	7,000	597.39	66.77	300	766	31.64
3	63	5,599	601.86	69.07	300	770	32.52
4	63	5,614	601.53	68.49	300	754	30.51
5	63	5,617	590.71	69.15	300	763	32.22

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. Therefore, the item statistics presented for the SJT are not comparable to those presented for the cognitive sections due to the different measurement models used.

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. For instance, if female and male candidates of the same ability level perform very differently on an item, then the item may be measuring something other than the ability of the candidates, possibly some aspect of the candidates that is related to gender. The principles of test fairness require that examinations undergo scrutiny to detect and remove items that behave in significantly different ways for different groups based solely on these types of demographic characteristics. In DIF, the terms “reference group” and “focal group” are used for group comparisons and generally refer to the *majority* and the *minority* demographic groupings of the exam population, respectively.

Detection of DIF

There are a number of procedures that can be used to detect DIF. One of the most frequently used is the Mantel-Haenszel procedure (Zwick, Thayer, Lewis, 1999). The Mantel-Haenszel procedure 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.

The Mantel-Haenszel procedure requires a criterion of proficiency or ability that can be used to match (group) candidates to various levels of ability. For the UCAT examination, matching is carried out using the raw score on each subtest associated with the item under study.

Items were classified for DIF using the Mantel-Haenszel delta statistic. This DIF statistic (hereafter known as MH D-DIF) is expressed as *differences* on the delta scale, which is commonly used to indicate the difficulty of test items. For example, an MH D-DIF value of 1.00 means that one of the two groups being analysed found the question to be one delta point more difficult than did *comparable* members of the other group. (Except for extremely difficult or easy items, a difference of one delta point is approximately equal to a difference of 10 points in percent correct between groups.) The convention of having negative values of MH D-DIF reflect an item that is differentially more difficult for the focal group (generally, females or the ethnic minority group) has been adopted. Positive values of MH D-DIF indicate the item is differentially more difficult for the reference group (generally white or male candidates). Both positive and negative values of the DIF statistic are found and are taken into account by these procedures.

Criteria for Flagging Items

For the UCAT examination, MH D-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. 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

The scaled units are based on a delta transformation of the proportion-correct measure of item difficulty. The delta for an item is defined as delta = 4z + 13 where z is the z-score that cuts off *p* (the proportion correct for an item) in the standard normal distribution. The delta scale removes some of the non-linearity of the proportion correct scale and allows easier interpretation of classical item difficulties.

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.

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 and social-economic status. Age was separated into groups 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 are 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 (e.g., between White and Black, Asian, Chinese, and Mixed race).

DIF Results Cognitive Subtests

Table 17 (operational items) and Table 18 (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 the operational pools, there were 14 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 14 occurrences, one occurred in the Age <20 / >35 comparison; two in the Male/Female comparison; seven in the White/Black comparison; and four in the White/Chinese comparison. For the pretest items, there was two occurrences of Category C DIF in the Male/Female comparison group. 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 19 (operational items) and Table 20 (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 N<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, White and Mixed, and between the NS-SEC Classes 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 was one occurrence of Category C DIF in the UK/Non-UK comparison, and 30 instances of Category B DIF overall. For the pretest items, there were two 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: DIF Summary Tables

Table 17. DIF Classification: Operational Pool

Comparison Group	MH D-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Making	
		N Items	Percentage	*N* Items	Percentage	*N* Items	Percentage	*N* Items	Percentage
Male/Female	A	196	98%	160	100%	249	100%	129	99%
	B	3	2%	0	0%	1	0%	0	0%
	C	1	0%	0	0%	0	0%	1	1%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	160	100%	250	100%	130	100%
Age <20/>35	A	28	14%	27	17%	43	17%	21	16%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	1	1%
	NA	172	86%	133	83%	207	83%	108	83%
	Total	200	100%	160	100%	250	100%	130	100%
White/Black	A	195	97%	157	98%	247	99%	124	95%
	B	2	1%	0	0%	3	1%	5	4%
	C	3	2%	3	2%	0	0%	1	1%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	160	100%	250	100%	130	100%
White/Asian	A	198	99%	160	100%	249	100%	125	96%
	B	2	1%	0	0%	1	0%	5	4%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	160	100%	250	100%	130	100%
White/Chinese	A	200	100%	159	99%	250	99%	126	97%
	B	0	0%	0	0%	0	0%	1	1%
	C	0	0%	1	1%	0	0%	3	2%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	200	100%	160	100%	250	100%	130	100%
White/Mixed	A	200	100%	159	99%	250	100%	130	100%
	B	0	0%	1	1%	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%	160	100%	250	100%	130	100%
NS-SEC Class 1/2	A	200	100%	160	100%	250	100%	129	99%
	B	0	0%	0	0%	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%	160	100%	250	100%	130	100%
NS-SEC Class 1/3	A	200	100%	160	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%	160	100%	250	100%	130	100%
	A	200	100%	160	100%	250	100%	129	99%

Comparison Group	MH D-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Making	
		N Items	Percentage	*N* Items	Percentage	*N* Items	Percentage	*N* Items	Percentage
NS-SEC Class 1/4	B	0	0%	0	0%	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%	160	100%	250	100%	130	100%
NS-SEC Class 1/5	A	199	100%	160	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%	160	100%	250	100%	130	100%

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

Table 18. DIF Classification: Pretest Pool

Comparison Group	MH D-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Making	
		N Items	Percentage	*N* Items	Percentage	*N* Items	Percentage	*N* Items	Percentage
Male/Female	A	240	100%	217	100%	260	100%	245	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	1	0%	0	0%	1	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	240	100%	218	100%	260	100%	246	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	240	100%	218	100%	260	100%	246	100%
	Total	240	100%	218	100%	260	100%	246	100%
White/Black	A	7	3%	27	12%	74	28%	1	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	233	97%	191	88%	186	72%	245	100%
	Total	240	100%	218	100%	260	100%	246	100%
White/Asian	A	240	100%	218	100%	260	100%	142	58%
	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%	104	42%
	Total	240	100%	218	100%	260	100%	246	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	240	100%	218	100%	260	100%	246	100%
	Total	240	100%	218	100%	260	100%	246	100%
White/Mixed	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%

Comparison Group	MH D-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Making	
		N Items	Percentage	*N* Items	Percentage	*N* Items	Percentage	*N* Items	Percentage
	C	0	0%	0	0%	0	0%	0	0%
	NA	240	100%	218	100%	260	100%	246	100%
	Total	240	100%	218	100%	260	100%	246	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	240	100%	218	100%	260	100%	246	100%
	Total	240	100%	218	100%	260	100%	246	100%
NS-SEC Class 1/3	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	240	100%	218	100%	260	100%	246	100%
	Total	240	100%	218	100%	260	100%	246	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	240	100%	218	100%	260	100%	246	100%
	Total	240	100%	218	100%	260	100%	246	100%
NS-SEC Class 1/5	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	240	100%	218	100%	260	100%	246	100%
	Total	240	100%	218	100%	260	100%	246	100%

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

Table 19. SJT DIF Classification: Operational Pool

Comparison Group	Degree of DIF					
	A		B		C	
	N Items	%	*N* Items	%	*N* Items	%
Male/Female	188	99%	1	1%	0	0%
Age <20/>35	185	98%	4	2%	0	0%
White/Black	183	97%	6	3%	0	0%
White/Asian	181	96%	8	4%	0	0%
White/Chinese	188	99%	1	1%	0	0%
White/Mixed	189	100%	0	0%	0	0%
NS-SEC Class 1/2	189	100%	0	0%	0	0%
NS-SEC Class 1/3	189	100%	0	0%	0	0%
NS-SEC Class 1/4	189	100%	0	0%	0	0%
NS-SEC Class 1/5	189	100%	0	0%	0	0%
UK/Non-UK	181	96%	7	4%	1	1%
English First Language/Other First Language	186	98%	3	2%	0	0%
Graduate/Non-Graduate	189	100%	0	0%	0	0%

Table 20. 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	235	96%	11	4%	0	0%	0	0%
Age <20/>35	236	96%	10	4%	0	0%	0	0%
White/Black	70	28%	5	2%	0	0%	171	70%
White/Asian	230	93%	14	6%	2	1%	0	0%
White/Chinese	28	11%	1	0%	0	0%	217	88%
White/Mixed	44	18%	1	0%	0	0%	201	82%
NS-SEC Class 1/2	218	89%	2	1%	0	0%	26	11%
NS-SEC Class 1/3	231	94%	2	1%	0	0%	13	5%
NS-SEC Class 1/4	202	82%	2	1%	0	0%	42	17%
NS-SEC Class 1/5	224	91%	4	2%	0	0%	18	7%
UK/Non-UK	235	96%	11	4%	0	0%	0	0%
English First Language/Other First Language	238	97%	8	3%	0	0%	0	0%
Graduate/Non-Graduate	237	96%	9	4%	0	0%	0	0%