University Clinical Aptitude Test (UCAT) Technical Report

Testing Interval: 26 July 2021 to 29 September 2021

Prepared by:
Pearson VUE
2 February 2022

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.

Table of Contents

Table of Tables

Table of Figures

Executive Summary

The University Clinical Aptitude Test (UCAT) was administered in 2021 from 26 July to 29 September. This report covers 37,217 exams that were delivered during that period, which is an increase of 9% on 2020. The exam was delivered in two modes: online and test centre. Online test delivery accounted for only 0.6% of candidates, so it is not possible to reliably compare results between these two groups.

Four versions of the UCAT were made available for candidates with special educational needs (SEN). Six percent of candidates who took the UCAT opted for a SEN version, and, similarly to previous years, candidates who took SEN versions of the exam outperformed those who took the non-SEN version.

Each exam consists of five subtests. Average scaled scores were stable for Verbal Reasoning, (VR), Quantitative Reasoning (QR) and Abstract Reasoning (AR), changing by just one or two scaled score points since 2020. Scores on the Decision Making (DM) subtest fell by 15 scaled score points, and the proportion of candidates falling into the lowest Situational Judgement Test (SJT) band increased from 9% in 2020 to 17% in 2021.

The 2021 UCAT consisted of five test forms. Reliabilities for the forms were good across the board and corresponding standard errors of measurement (*SEM*s) were satisfactorily low, and consistent with previous years.

The cognitive subtests were quite speeded, with the majority of candidates using all the available time and the average time used very close to the available time. Speededness reduced in the SEN exams where candidates have more time available. The SJT remains the least speeded subtest.

Demographic trends in 2021 were consistent with those of previous years, with higher scores being associated with higher socio-economic classification (SEC), Chinese or White ethnicity, speaking English as a first language, and being a UK resident. Males tended to perform better than females on the cognitive subtests, but females outperformed males on the SJT.

Individual item analysis showed satisfactory quality for the majority of operational items. Pretesting is intended to identify poor-quality items before they enter the operational scored test, and therefore the pretest items ranged more broadly in quality and on the whole performed less well. Five cognitive operational items and 75 cognitive pretest items failed the quality criteria and were removed from the item bank, whereas 20 SJT operational items and 172 SJT pretest items failed. Additionally, six operational items and eight pretest items were removed due to potentially exhibiting bias.

1. Introduction

The purpose of the UCAT 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. It is not an exam that measures student achievement and therefore it does not contain any curriculum or science content.

This report covers the 2021 UCAT that was delivered from 26 July 2021 to 29 September 2021. The design of the exam has remained the same as in recent years. As outlined in Section 2, it consisted of five subtests ranging from 26 to 69 items each. Section 3 describes the exam results in terms of candidate volumes, scaled scores, and SJT bands. Section 3 reports exam results in reference to candidates who qualified for a SEN version of the exam, whether candidates applied for medicine or dentistry, the mode of delivery, and candidate demographic characteristics.

Following the analysis of results by demographic, exam timing is examined in Section 4. Section 5 deals with analysis of the five test forms, Section 6 summarises analysis of the test items, and the final section of this report provides recommendations.

2. Exam Design 2021

The 2021 UCAT consisted of five balanced test forms each with five subtests. Each subtest includes scored and unscored items as shown in Table 1 below.

Table 1. UCAT Exam Design

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

Candidates were allowed 120 minutes to answer a total of 233 items from the five subtests. There were four groups of candidates with extra time allowances in 2021. The timing and scoring of the SEN exams are explored in detail in section 3.2 below.

Raw scores in each cognitive subtest were transformed to a scaled score ranging from 300 to 900. SJT scaled scores ranged from 300 to 834. Universities received cognitive subtest scaled scores plus a total score; a simple sum of the four cognitive subtest scores ranging from 1,200 to 3,600. SJT scaled scores are further categorised into four bands. The bands are determined by scaled score ranges as defined in Table 2 below.

Table 2. SJT Band Scaled Score Range and Description

Bands	Scaled Score Range	Intended Band Proportions	Narrative
Band 1	671–900	21%	Those in Band 1 demonstrated an excellent level of performance, showing similar judgement in most cases to the panel of experts.
Band 2	611–670	38%	Those in Band 2 demonstrated a good, solid level of performance, showing appropriate judgement frequently, with many responses matching model answers.
Band 3	530–610	31%	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	300–529	10%	The performance of those in Band 4 was low, with judgement tending to differ substantially from ideal responses in many cases.

The 2021 UCAT was delivered in two modes: the OnVUE mode, where a candidate can take the test in their own home with an online proctor, or the test centre mode, where candidates take the test in a specially designed test centre. Only 231 candidates took the online version of the test (see 3.4 below).

3. Examination Results

3.1 Overall Exam Results

This report covers examination results for 37,217 candidates who took the UCAT during the period 26 July 2021 to 29 September 2021. Candidate volumes have increased each year since 2017, and increased by 9% since 2020, as illustrated in Figure 1 below.

Figure 1. Candidate Volumes since 2017

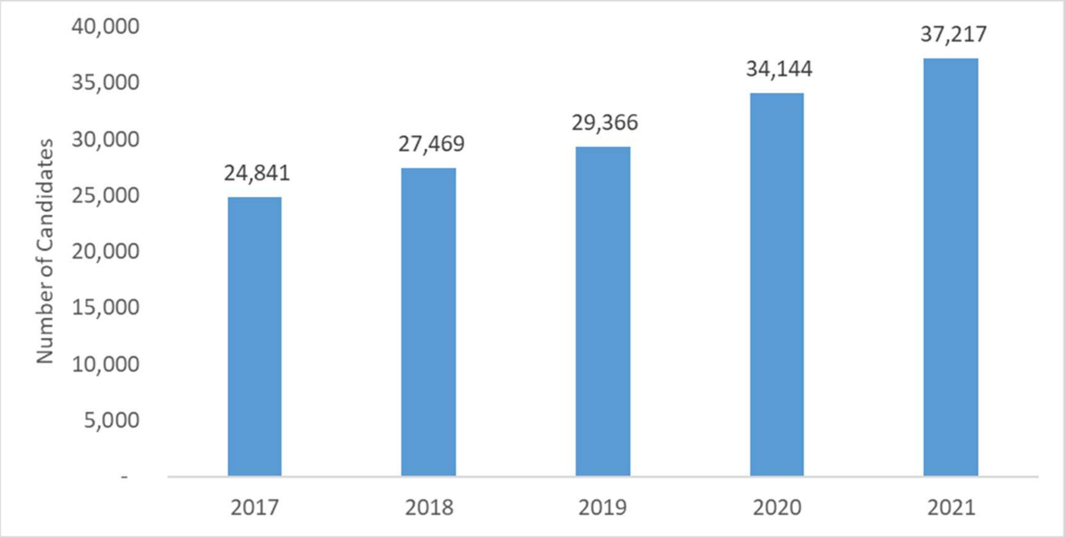


Table 3 presents summary statistics for each of the cognitive subtests plus the total scaled score for the cognitive subtests. VR scores were lowest with a mean score of 572, the highest average score was achieved on QR with an average of 665.

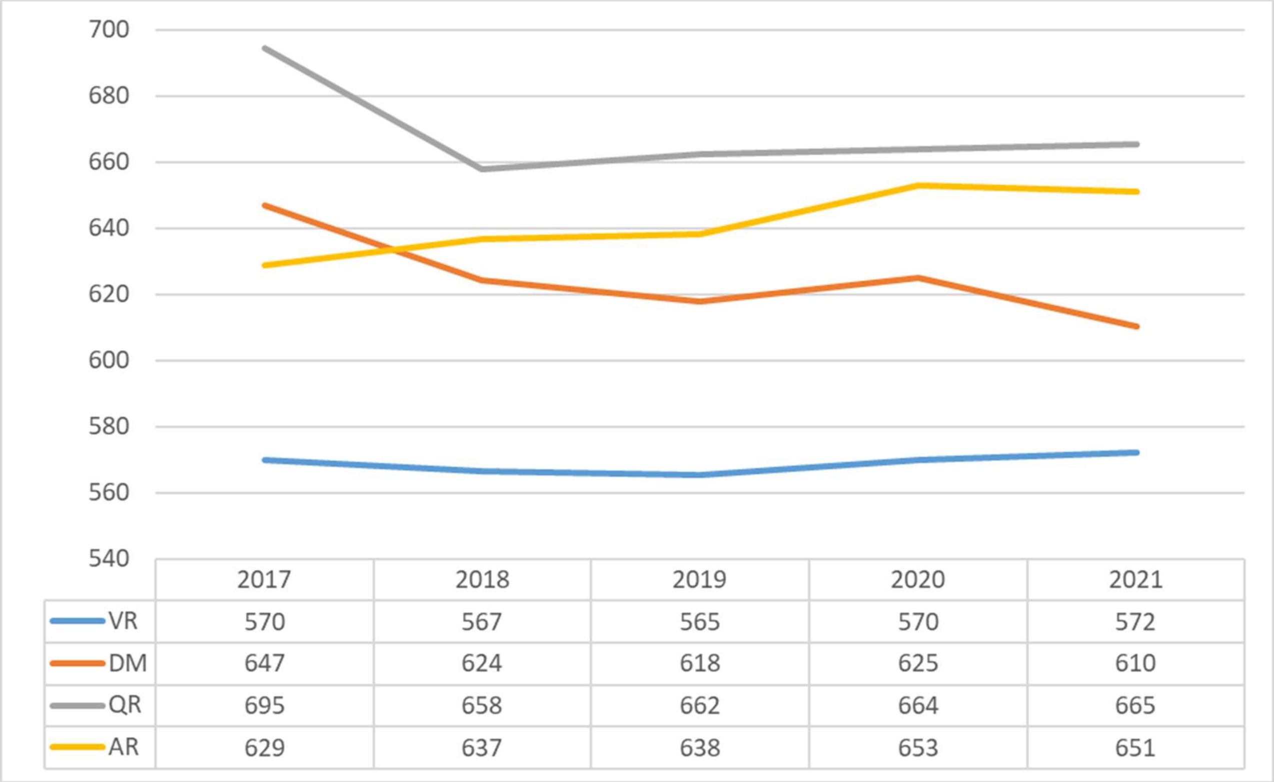
Table 3. Cognitive Subtest and Total Scaled Score Summary Statistics

Subtest	Mean	SD	Min	Max
VR	572.14	75.15	300	900
DM	610.09	88.52	300	900
QR	665.42	79.83	330	900
AR	651.2	94.33	300	900
Total	2,498.84	275.23	1,380	3,500

Figure 2 shows the change in scaled scores since 2017. The year 2017 was chosen as a start point for comparison because prior to 2017 there was no operational DM section.

Over the five-year period, QR and DM have tended to fall. The large drops in 2018 were associated with a change to the scaling method for QR and a change in the benchmark population for DM. Both changes were intended to bring the scaled scores closer to 600. Since 2017, AR has tended to rise and VR has remained stable.

Figure 2. Scaled Scores by Year since 2017

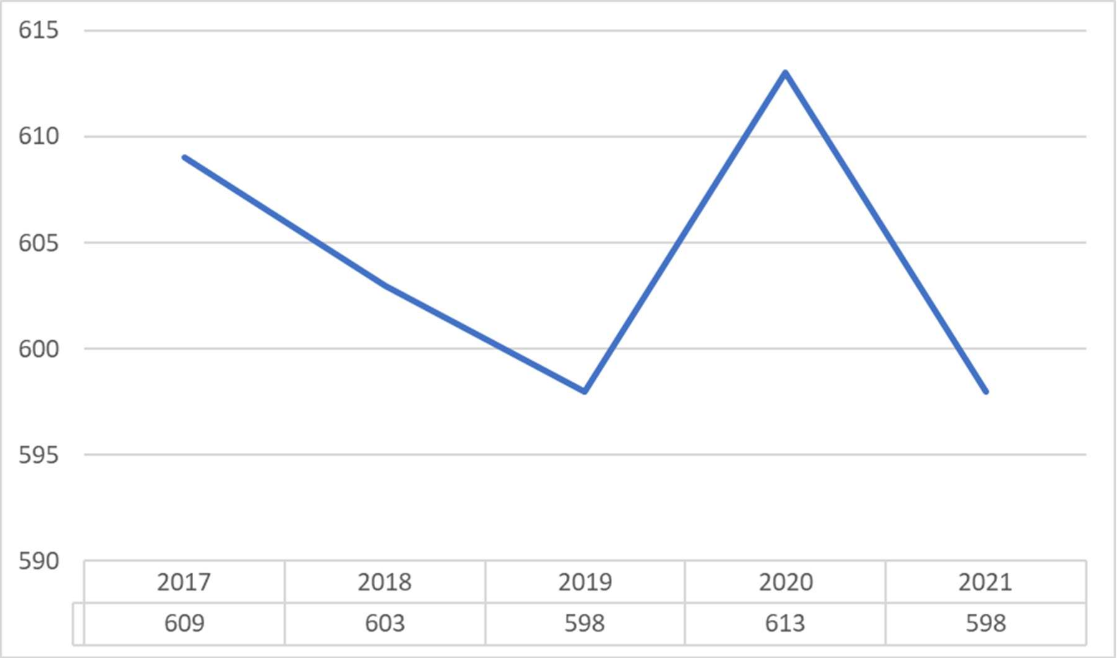


	2017	2018	2019	2020	2021
VR	570	567	565	570	572
DM	647	624	618	625	610
QR	695	658	662	664	665
AR	629	637	638	653	651

Since 2020 VR, QR, and AR have remained stable, changing by one or two scaled score points. DM, by contrast, has fallen by 15 scaled score points. This fall in scores is within one *SEM*, which ranges from 43.4 to 43.9 for DM (as discussed in Section 5). This means that statistically the effect is not large enough to warrant concern. It is not clear why DM scores dropped when the other subtests tended to remain stable, neither demographic differences nor timing appear to account for it (as explored in Sections 3.5 and 4 respectively).

For the SJT, average scale score was 598, and the standard deviation of scores was 75. Figure 3 shows the average SJT scaled score has ranged between 598 and 613 since 2017.

Figure 3. SJT Scaled Scores 2017-2021



The number and percentage of candidates in each band for the 37,217 candidates who took the 2021 UCAT are shown in Table 4. The target figure presented in the Table 4 is the anticipated number of candidates who should fall into each band given the band boundaries and the previous year's distribution of scores.

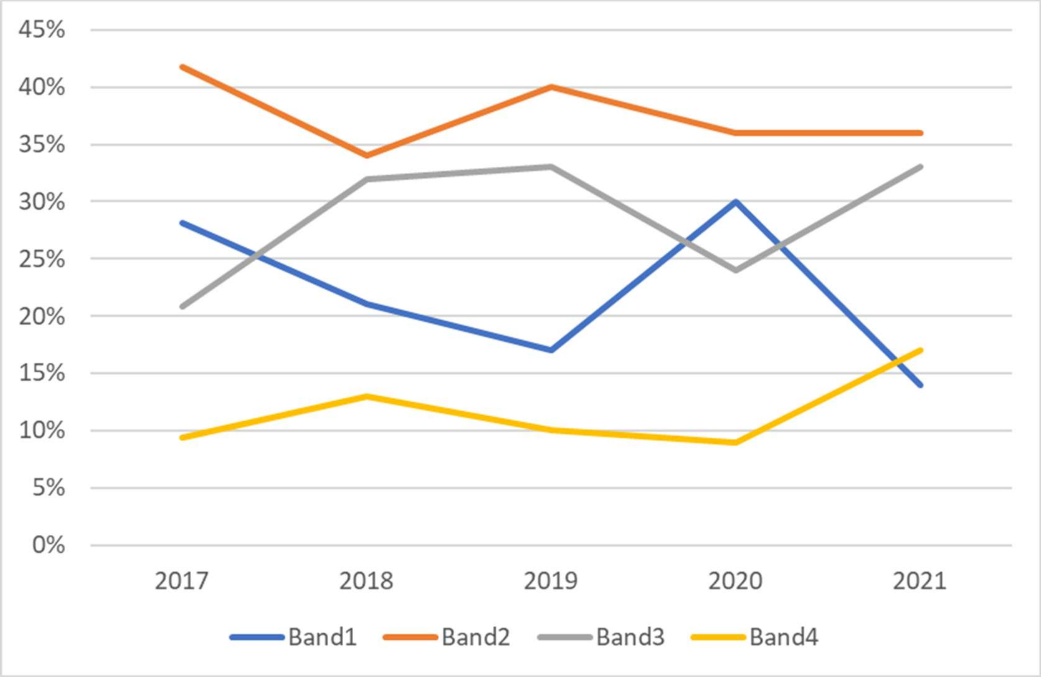
Table 4. SJT Band Distribution in 2021

SJT Band	Number of Candidates	Percentage of Candidates	Mean Scaled Score	Target %
Band 1	5,189	14%	691	21%
Band 2	13,552	36%	640	38%
Band 3	12,335	33%	576	31%
Band 4	6,141	17%	469	10%
Total	37,217	100%	598	100%

The proportion of candidates falling into Bands 1 and 4 deviates from the target. Candidates categorised as Band 1 are seven percentage points lower than the target, and candidates categorised as Band 4 are seven percentage points higher.

Each year the target proportion can change. Figure 4 illustrates the distribution of candidates across SJT bands since 2017.

Figure 4. SJT Band Proportions 2017–2021



The equating method undertaken when constructing test forms ensures that the difficulty of the test forms is controlled year-on-year, meaning test construction is not the source of the shifts in performance we see in Figure 3. It is more likely that the difference is related to 2021 candidates being less able in the traits measured by the SJT than candidates who took the test in 2020. It may also be the case that the distribution of candidate scores is influential.

The distribution of scores is important because the band boundaries, defined in Table 2 above, are set each year in reference to candidate performance in the prior year. Candidate performance in 2020 was relatively high, with an increase in candidates being categorised as Band 1. This increase resulted in the boundary for Band 1 being higher in 2021 than in 2020; therefore, when candidate performance fell in 2021, correspondingly fewer candidates were categorised as Band 1. In short as the band boundaries increased from 2020 to 2021, the proportion of candidates achieving high scaled scores fell.

3.2 Special Educational Needs

There are four exams available for SEN candidates who are allowed extra time and breaks. Time allowances for each subtest and exam are illustrated below in Table 5.

Table 5. Exam Version Time Allowed

Subtest	UCAT	UCATSEN	UCATSENSA	UCATSEN50	UCATSA
VR	00:21:00	00:26:15	00:26:15	00:31:30	00:21:00
DM	00:31:00	00:38:45	00:38:45	00:46:30	00:31:00
QR	00:24:00	00:30:00	00:30:00	00:36:00	00:24:00
AR	00:13:00	00:16:15	00:16:15	00:19:30	00:13:00
SJT	00:26:00	00:32:30	00:32:30	00:39:00	00:26:00

Only 6% of candidates took a SEN version of the exam with the most popular being the UCATSEN as shown in Table 6.

Table 6. Exam Version Candidate Volumes

Exam	N	%
UCAT	34,841	94%
UCATSEN	1,904	5%
UCATSENSA	248	1%
UCATSEN50	135	0%
UCATSA	89	0%
Total	37,217	100%

Historically candidates who take a SEN version of the exam usually outperform candidates who take the non-SEN version. Table 7 summarises the scaled score statistics by exam version. SEN candidates outperformed non-SEN candidates in all four subtests. The sample size of UCATSEN50, UCATSA, and UCATSENSA are small and results for those versions should be treated with caution.

Table 7. SEN and Non-SEN Cognitive Subtest

Subtest	Statistic	UCAT (34,841)	UCATSEN (1,904)	UCATSENSA (248)	UCATSEN50 (135)	UCATSA (89)
VR	Mean	570.55	593.43	611.53	605.7	575.62
	SD	74.63	78.46	79.66	82.66	71.67
	Min	300	300	420	400	460
	Max	900	890	830	830	760
DM	Mean	608.8	625.19	652.54	632.96	640.9
	SD	88.01	93.18	95.2	95.42	85.77
	Min	300	300	390	360	440
	Max	890	890	880	900	890
QR	Mean	664.52	673.72	700.89	698.67	687.75
	SD	79.6	80.21	87.22	87.24	81.92
	Min	330	430	460	500	500
	Max	900	900	900	900	900
AR	Mean	649.92	668.25	673.67	691.04	664.04
	SD	94.27	92.36	96.29	102.7	85.3
	Min	300	330	320	470	420

	Max	900	900	890	890	880
Total	Mean	2,493.80	2,560.60	2,638.63	2,628.37	2,568.32
	SD	274.14	277.32	289.94	302.90	253.74
	Min	1,380	1,460	1,890	1,840	1,950
	Max	3,500	3,350	3,410	3,370	3,290

Table 7 also includes average total cognitive score for each exam version. Clearly SEN candidates performed better than non-SEN candidates on the cognitive subtests as a whole. The difference between candidates who sat the UCAT and those who sat the UCATSEN amounts to 67 scaled score points. As Table 8 shows the difference in scores between the non-SEN version of the exam and the UCATSEN exam has fallen since 2017.

Table 8. SEN and Non-SEN Historical Score Difference

Year	Difference in score between non-SEN and the most popular SEN exam
2021	67
2020	77
2019	82
2018	103
2017	79

The pattern of SEN candidates being stronger than non-SEN candidates is repeated for the SJT results, where the UCAT version of the exam has the lowest proportion of candidates in Band 1 and the highest in Band 4. The breakdown of SJT band proportions and scaled scores by exam version is presented in Table 9 below. The version of the exam on which candidates performed the best is UCATSENSA, where 28% of candidates are categorised as Band 1 and only 2% are categorised as Band 4, but note the prior warning that few candidates sat that version of the exam, meaning comparison may not be reliable.

Table 9. SJT Band by Exam Version

		% of Candidates			
Exam Version	Mean Scaled Score	Band 1	Band 2	Band 3	Band 4
UCAT	596	13.50%	36.06%	33.48%	16.97%
UCATSEN	615	19.28%	40.76%	28.83%	11.13%
UCATSENSA	638	27.82%	44.76%	25.00%	2.42%
UCATSEN50	631	25.93%	42.22%	26.67%	5.19%

One potential reason for SEN candidates outperforming non-SEN candidates is the extra time they receive. After the 2020 exam, Pearson VUE undertook analysis to understand whether some of this difference may also be due to demographic differences between the SEN and non-SEN candidate groups. We matched 100 stratified samples of UCATSEN candidates to the demographic makeup of the UCAT candidates according to first language, gender, residency, age group, education level and SEC. The comparison of average scaled scores of the stratified sample of UCATSEN candidates to the UCAT candidates is shown in Table 10 below. We anticipated that when the samples were matched demographically, the UCATSEN scores would come closer to the UCAT results, and that is the case for the VR and DM subtests, as well as the total score. However, for QR, the average score did not change and for AR, it increased.

Table 10. Stratified Sample of 2020 UCAT

Subtest	UCAT 2020	UCATSEN Before/After Sampling	Difference Between UCAT/SEN Before/After Sampling
VR	569	From 587 to 579	From 18 to 10
DM	624	From 640 to 636	From 16 to 12
QR	663	From 683 to 683	From 20 to 20
AR	652	From 672 to 674	From 20 to 22
Total	2,508	From 2,582 to 2,572	From 74 to 64

In summary, it appears that some of the score differences we observed in the 2020 exam between the SEN and non-SEN versions of the test are the result of the demographic characteristics of the candidates who qualify for SEN exams. However, score differences between the versions do remain, and, in the case of AR, increased after sampling. It is likely that these differences are caused by a demographic difference that we do not currently measure and/or the extra time allocation.

The interaction between specific demographic variables and the SEN version of the exam has also been explored, and is discussed below in Section 3.6.

3.3 Medicine and Dentistry

Many candidates who take the UCAT also apply for medical or dental school via the Universities and Colleges Admissions Service (UCAS). This section of the report concerns the performance of candidates in relation to whether they apply to study medicine or dentistry.

The majority of candidates applied for medicine, accounting for 68% of candidates. Nine percent of candidates applied for dentistry, and the remaining 23% applied for neither.

Candidates who applied for medicine as a first choice outperformed those who applied for dentistry, as illustrated in Table 11. The highest average scaled score was achieved on QR and the lowest on VR for both candidate groups.

Table 11. Medicine and Dentistry Candidates: Cognitive and Total Scaled Scores

	Mean		SD	
Subtest	Medicine	Dentistry	Medicine	Dentistry
VR	586.8	559.89	73.89	64.66
DM	630.47	603.3	83.82	78.45
QR	682.6	664.27	77.65	73.37
AR	670.27	657.49	90.77	89.5
Total	2,570.14	2,484.95	258.09	241.88

Better performance by medicine candidates is also reflected in the SJT banding. As Table 11 shows, more medicine than dentistry candidates appeared in Band 1, and fewer medicine than dentistry candidates appeared in Band 4.

Table 12. Medicine and Dentistry Candidates: SJT Bands

		Band 1	Band 2	Band 3	Band 4
Group	Mean Scaled Score	% of Candidates			
Dentistry	602	13%	38%	35%	13%
Medicine	614	17%	42%	32%	9%

In summary UCAT candidates who apply for medicine perform better across all subtests than those who applied for dentistry. This is consistent with test performance in previous years.

3.4 Mode of Delivery

In 2021 the UCAT was offered in both the standard test centre and online proctored mode. Two hundred and thirty-one candidates took the exam in the online proctored mode, amounting to only 0.6% of all candidates, the other candidates all took the test in the standard test centre mode. This contrasts with 2020, when 32% of candidates took the exam in the online mode, and 68% in the test centre mode.

Given the large difference in volumes between the two modes and the low number of candidates who took the test in the online mode in 2021, it is not possible to draw reliable inferences on differences in performance for the 2021 cohort of candidates.

3.5 Examination Results by Demographic Variables

3.5.1 Variation by Demographic Group

Pearson VUE undertakes several tasks as part of the item development and analysis process to ensure differential performance related to demographic characteristics are not caused by the test content or mode of delivery. All content creators and reviewers complete an editorial course and agree to a global set of principles and best practices that need to be considered when creating content. Item writers and editors are provided with specific guidelines to be adhered to when creating content. Test items are developed using a group of content creation specialists, and bias, sensitivity, and accessibility reviews are undertaken before test items are used in the exam. We also produce practice resources, freely accessible to all candidates. Finally, we analyse the performance of individual items by demographic characteristic and removed any items that might exhibit bias (as discussed in 6.3 below).

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 subtests scaled scores.

3.5.2 Gender

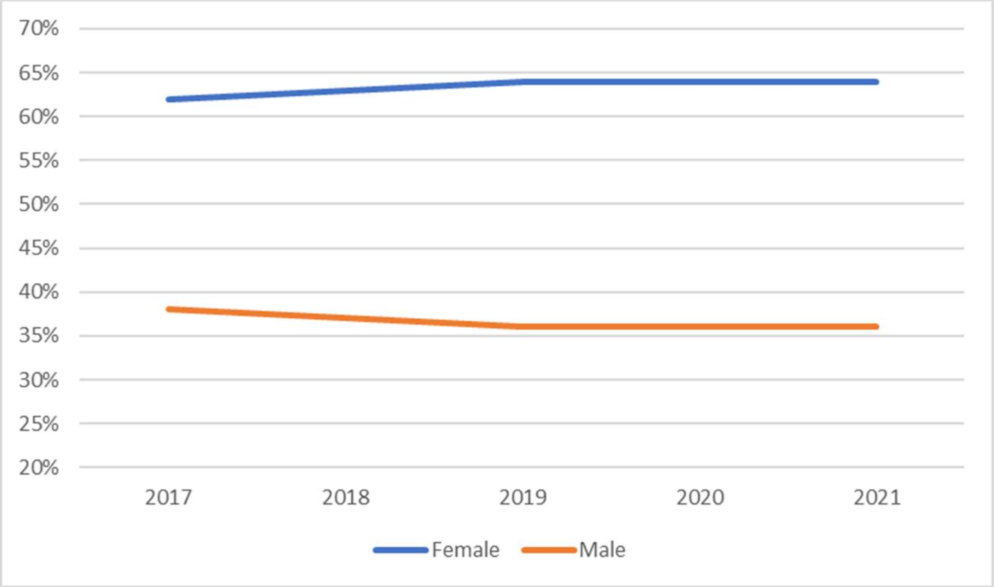
Table 13 presents the breakdown of test-takers by gender. The majority of test-takers were female, and only 181 stated "Other", or that they would prefer not to say.

Table 13. Gender Counts

Gender	*N*	%
Female	23,650	64%
Male	13,386	36%
I prefer not to say	131	0%
Other	50	0%

The distribution of candidates by gender has remained stable since 2017, with a slight increase in female candidates from 2017 to 2019.

Figure 5. Distribution of Candidates by Gender 2017–2021



Males outperformed females on all subtests except the SJT, where females performed better than males. The difference between male and female average scores is shown in Table 14 ranged from 11 scaled score points on VR, to 24 scaled score points on QR.

Table 14. Gender Scaled Scores

Subtest	Mean Scaled Score		*SD* Scaled Score	
	Female	Male	Female	Male
VR	567.92	579.11	74.32	75.93
DM	604.24	620.09	88.13	88.27
QR	656.54	680.96	77.71	81.05
AR	647.05	658.59	92.61	96.94
Total Cognitive	2,475.76	2,538.75	271.45	277.15
SJT	602.82	588.52	72.49	77.68

A statistical test was used to examine whether the differences between the two groups observed in Table 15 were statistically significant. Table 16 shows the *T*-statistic, degrees of freedom and *p* value for each subtest and the total cognitive scores. The *df* column shows the available sample size. A non-zero *T*-statistic indicates there is a difference in the average score between two group samples. However, the difference may or may not be statistically significant. That is, the difference may or may not be sufficient evidence of a true difference in the entire population (e.g., between all eligible males and all eligible females). The *p* value shows the probability due to chance of observing a particular *T*-statistic (or something more extreme). Lower *p* values (e.g., less than 0.01) indicate that we would be unlikely to see such a difference in our sample if there were no true difference in the population.

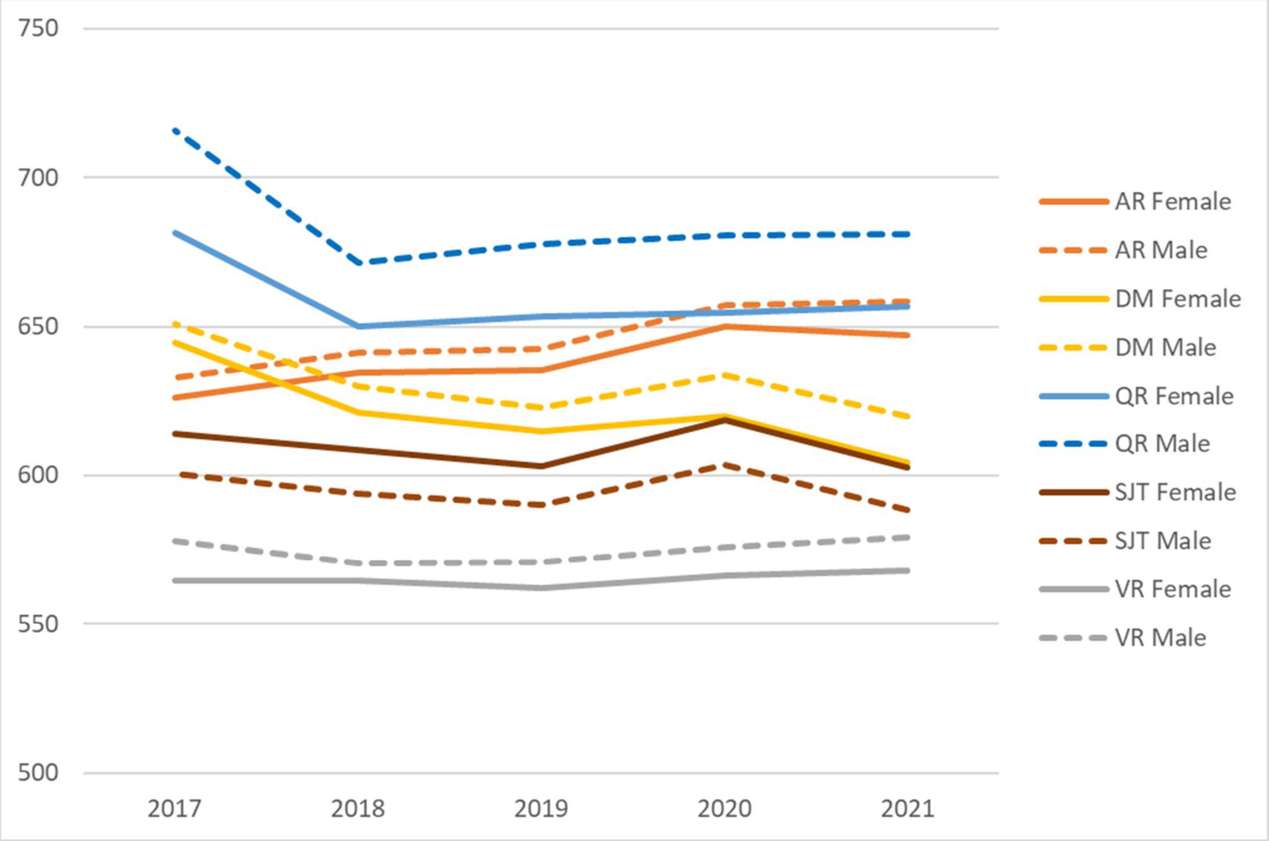
Therefore Table 15 shows that there are differences between male and female performance on each subtest and on the total cognitive scores, and that these differences are likely not to be the result of random chance.

Table 15. Gender *T*-Test

Subtest	*T*-Statistic	*df*	*p* Value
VR	13.81	37,034	< 0.01
DM	16.62	37,034	< 0.01
QR	28.6	37,034	< 0.01
AR	11.32	37,034	< 0.01
Total Cognitive	21.29	37,034	< 0.01
SJT	-17.76	37,034	< 0.01

Figure 6 illustrates that the subtest differences by gender. Differences have tended to be consistent year on year. Since 2017, the difference in scores between males and females has slightly broadened in the DM subtest, and since 2020 the range has slightly broadened in the AR subtest.

Figure 6. Distribution of Candidates by Gender 2017–2021



3.5.3 Ethnicity

UCAT candidates who reside in the UK are requested to answer a question relating to their ethnicity. During analysis we categorise these similarly to the categories used by the Office for National Statistics in the following way:

- UK – Asian: Asian Indian, Asian Pakistani, Asian Bangladeshi, Asian Other
- UK – White: White
- UK – Black: Black Caribbean; Black African, Black Other
- UK – Other: Other, e.g. gypsy, traveller, Irish traveller, or not specified.
- UK – Mixed Race: Mixed White and Black Caribbean, Mixed White and Black African, Mixed White and Asian, Other Mixed
- UK – Chinese

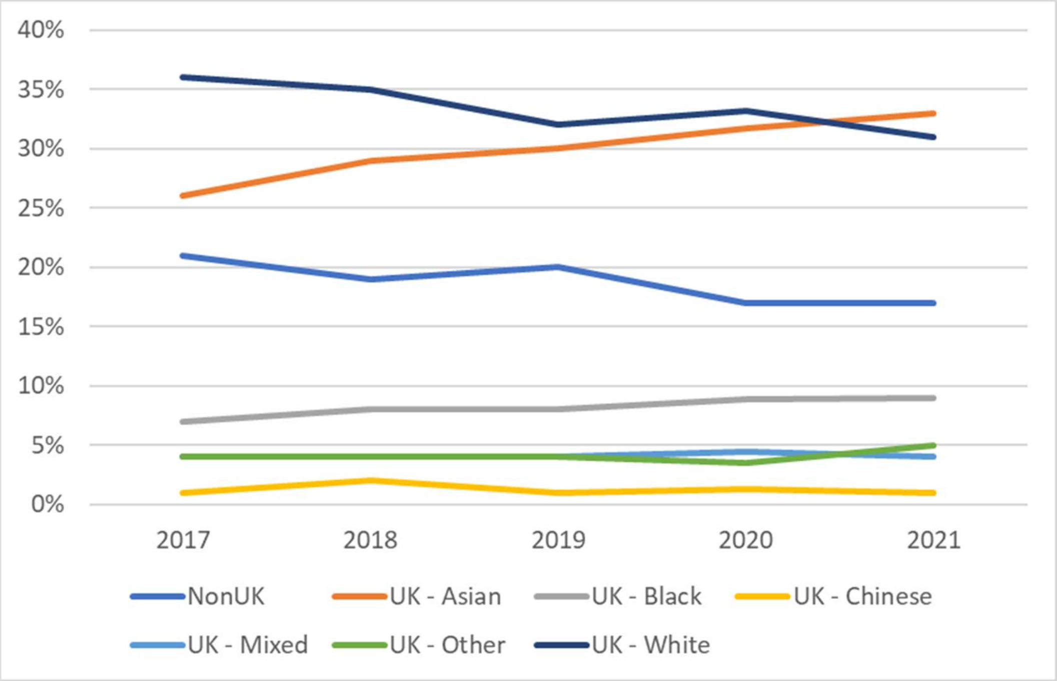
Table 16 shows the breakdown of candidates by ethnicity in the 2021 exam. The biggest candidate group was UK – Asian. Seventeen percent of candidates were not categorised due to being non-UK candidates.

Table 16. Ethnic Group

Country	Ethnic Group	*N*	Percent UK Candidates	Percent Total Candidates
UK	Asian	12,300	40%	33%
UK	White	11,616	37%	31%
UK	Black	3,321	11%	9%
UK	Other ethnic group	1,714	6%	5%
UK	Mixed	1,596	5%	4%
UK	Chinese	469	2%	1%
Non-UK	Non-UK	6,201		17%

The proportion of candidates in each ethnic group has remained fairly stable in recent years. Figure 7 shows that the most common ethnic group changed from White to Asian in 2021. The proportion of non-UK candidates has decreased since 2017 and the proportion of Black candidates has slightly increased.

Figure 7. Distribution of Candidates by Ethnic Group 2017–2021



UK – White candidates performed better on average on all subtests than other groups, except QR and AR where UK – Chinese candidates on average were the best performers. Table 17 shows the average scores in each subtest for each ethnic group. Each row is shaded from light to dark reflecting the range in scores from low to high, respectively. Performance was lowest for UK – Black candidates on average on all subtests except the SJT, where non-UK candidates received the lowest average scaled scores.

Table 17. Ethnic Group Mean Scaled Score

Subtest	Asian	White	Black	Other Ethnic Group	Mixed	Chinese	Non-UK
VR	564.2	596.22	549.2	550.66	583.02	588.4	556.95
DM	601.97	638.35	575.1	589.35	622.17	637.1	592.58
QR	668.45	677.89	629.69	650.34	670.93	713.56	654.27
AR	656.32	663.89	617.63	642.49	659.17	701.39	631.83
Total Cognitive	2,490.93	2,576.35	2,371.62	2,432.84	2,535.29	2,640.45	2,435.64
SJT	599.69	615.59	585.96	591.79	605.89	612.06	565.01

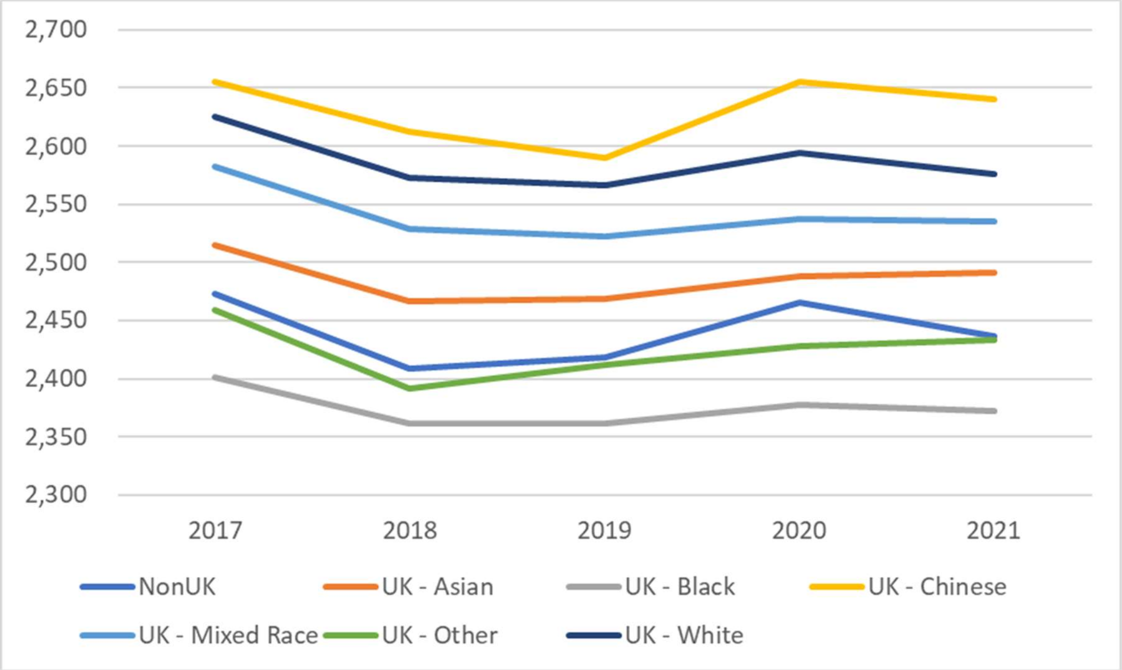
An *F*-test was used to examine whether the differences observed in Table 18 were likely to be due to chance. An *F*-test is similar to the *T*-test discussed in relation to gender (see 3.5.2). It is used when there are more than two groups. Table 18 has a positive *F*-statistic for each subtest and a *p* value of less than 0.01, which indicates that the differences observed in Table 17 above are likely to reflect true differences in performance in the candidate population.

Table 18. Ethnic Group *F*-test

Subtest	*F*-Statistic	*df*	*p* Value
VR	368.58	6	< 0.01
DM	392.85	6	< 0.01
QR	229.24	6	< 0.01
AR	186.59	6	< 0.01
Total Cognitive	393.15	6	< 0.01
SJT	351.06	6	< 0.01

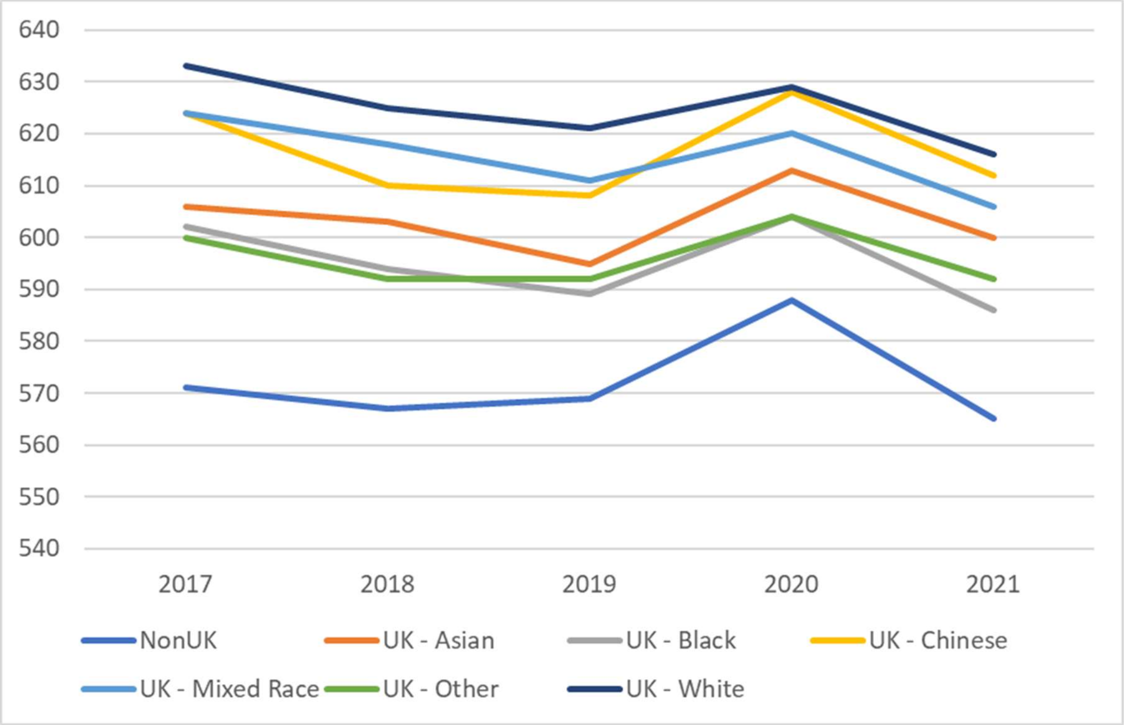
Average total cognitive scaled scores fell for all ethnic groups between 2017 and 2018 reflecting the rescaling that took place. After 2018 score remained fairly stable for UK White, UK Mixed Race and UK Black, with small increases for UK Asian and larger increases for UK Chinese, non UK and UK Other.

Figure 8. Ethnic Group Mean Total Scaled Score 2017–2021



In the SJT there was a fairly large increase in scores for all ethnic groups between 2019 and 2020 and a slightly larger fall for all groups between 2020 and 2021. The most notable ethnic group trend for the SJT is the margin by which non-UK candidates underperform relative to the other groups, as can be observed in Figure 9.

Figure 9. Ethnic Group Mean Scaled Score for SJT 2017–2021



The underperformance of non-UK candidates on the SJT might be explained by a link between situational judgement and cultural competence. Specifically, that UK based candidates are more likely to have better understanding of UK-specific situational norms of behaviour. However, no potential bias against candidates on the basis of residency was identified at the item level in the SJT (see 6.3.4, and the discussion in the Recommendations in Section 7).

3.5.4 Socio-Economic Classification

UK candidates are asked several questions relating to their parent's or carer's work to categorise them into SECs. These questions ask candidates to state what type of employment the parent or carer does, whether they are employed or self-employed, and the number of people they work with if employed, or employ if self-employed. Although the primary question about what work the parent or carer does is mandatory, if a candidate responds with "don't know", "prefer not to say" or "never worked" it is not possible to categorise them into an SEC. Therefore, we typically see a large proportion of UK candidates not being categorised into one of the five SECs.

This issue is illustrated in Table 19, which shows that 27% of all candidates reside in the UK but cannot be categorised into an SEC. The candidates who can be categorised fall predominantly into SEC 1, representing Managerial and Professional Occupations.

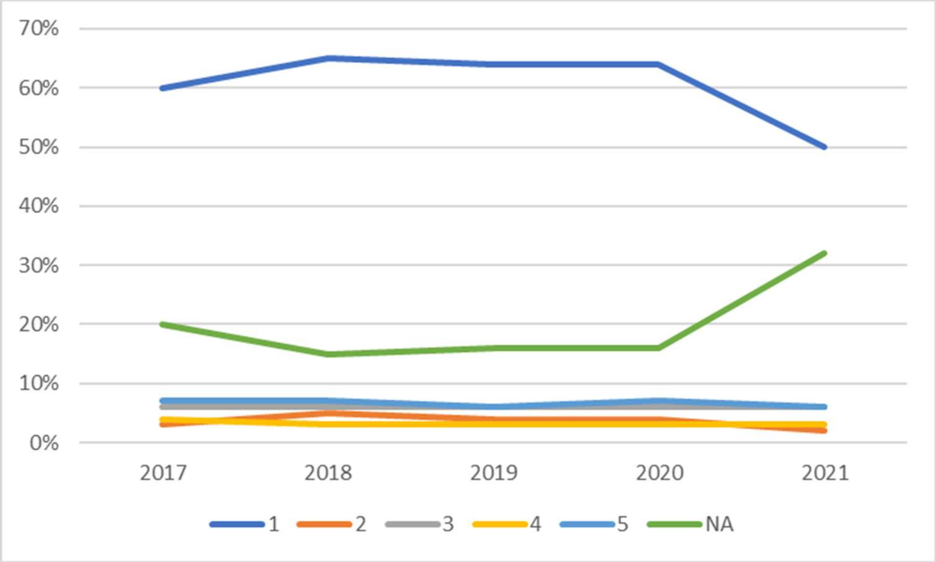
Table 19. SEC Counts

Country	SEC	*N*	% of SEC	% of All
UK	1	15,543	74%	42%
	2	644	3%	2%
	3	1,859	9%	5%
	4	1,077	5%	3%
	5	1,975	9%	5%
	Unknown	9,918		27%
EU		1,215		3%
Other		4,986		13%

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

Prior to 2021, SEC was calculated for up to two parents or carers, then candidates were categorised as the highest of the two SECs. However, in 2021 the SEC questions changed to ask candidates to enter responses for only the highest earning parent or carer. The result is that proportionally more candidates appear in the NA category in 2021 than in previous years, as illustrated in Figure 10. There were fewer candidates in SEC 1 in 2021 than in previous years; however, since this fall corresponds to a similar rise in SEC NA, it is likely that the new way of measuring SEC is influencing this measure.

Figure 10. Candidates by SEC 2017–2021



Candidates who are SEC 1 achieve higher scores than all other classifications, as shown in Table 20.

Table 20. Scaled Scores by SEC

Mean Scaled Score						
Subtest	SEC 1	SEC 2	SEC 3	SEC 4	SEC 5	NA
VR	587.54	581.43	566.41	564.53	558.6	561.47
DM	630.12	610.89	604.78	602.69	589.97	595.41
QR	679.86	659.46	659.39	659.39	652.71	654.45
AR	667.76	643.8	648.02	643.35	639.17	641.68
Total Cognitive	2,565.29	2,495.57	2,478.60	2,469.96	2,440.46	2,453.01
SJT	613.48	606.46	600.08	598.47	595	592.89
SD						
VR	74.3	73.55	67.85	69.98	69.13	72.53
DM	84.29	83.81	79.66	83.59	81.37	89.08
QR	77.22	76.36	73.21	73.09	73.66	77.71
AR	92.46	86.78	89.16	88.24	89.41	92.92
Total Cognitive	262.01	255.1	246.06	248.63	248.76	272.35
SJT	64.58	65.43	67.25	70	72.03	75.43

As with the other demographic categories, hypothesis testing was used to examine whether the scores are likely to be true reflections of the candidate population. Table 21 shows that the score differences observed in each subtest are likely to be due to true differences.

Table 21. SEC *F*-Test

Subtest	*F*-Statistic	df	*p* Value
VR	191.07	5	< 0.01
DM	244.33	5	< 0.01
QR	160.51	5	< 0.01
AR	120.67	5	< 0.01
Total Cognitive	268.54	5	< 0.01
SJT	119.22	5	< 0.01

3.5.5 Age

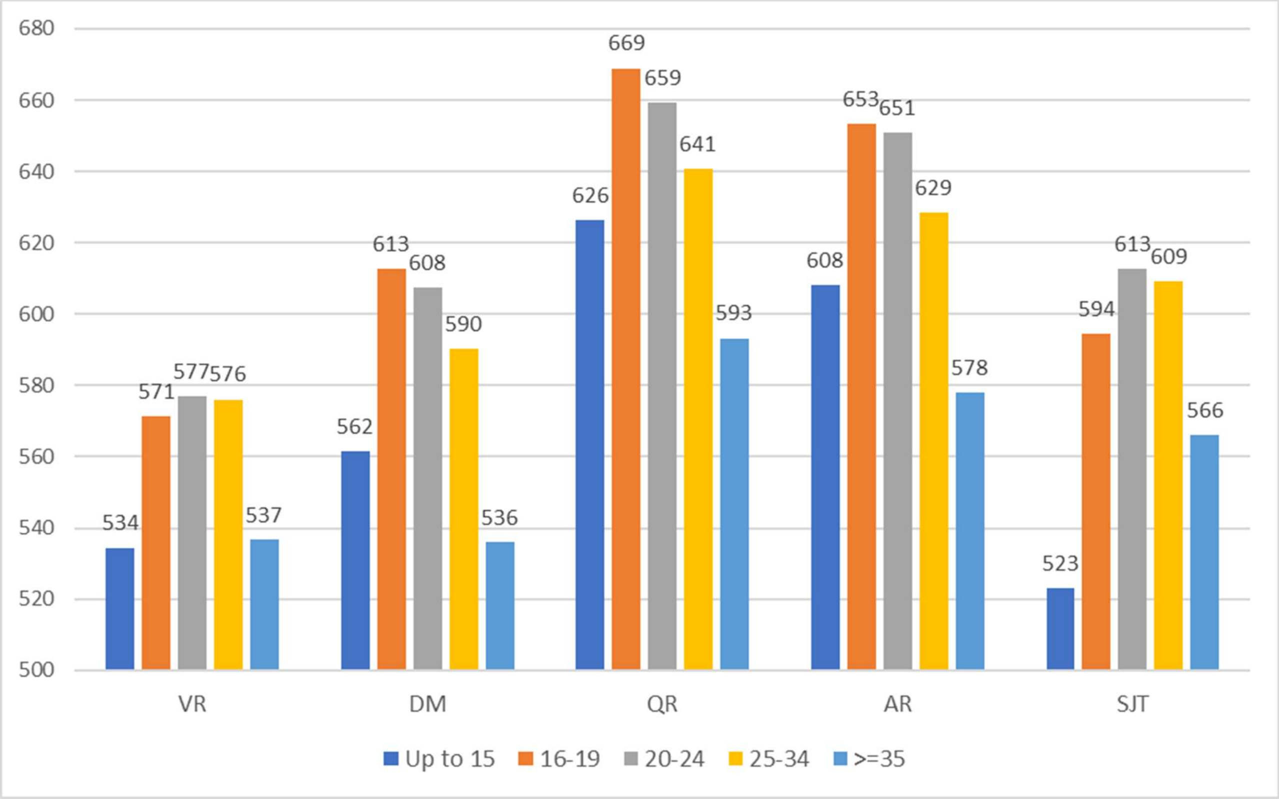
The majority of UCAT candidates are aged 16–19 years old. A small minority of candidates are 35 or older and an even smaller proportion are under 16.

Table 22. Age Counts

Age	*N*	Percent
<= 15	64	0%
16–19	29,172	78%
20–24	6,002	16%
25–34	1,678	5%
>= 35	301	1%

Candidates who were aged 16–19 tended to perform better in the DM, QR and AR subtests, as illustrated in Figure 11. In the SJT and VR, candidates who were 20–24 tended to perform the best. Candidates who were under 16 and over 34 typically had lowest performance on the exam; however, the small group sizes for those categories means it is difficult to draw meaningful conclusions from that information.

Figure 11. Average Scaled Scores by Age



Hypothesis testing demonstrated that the differences observed among the groups is unlikely to have occurred due to chance, as shown in Table 23.

Table 23. Age *F*-Test

Subtest	*F*-Statistic	df	*p* Value
VR	29.2	4	< 0.01
DM	87.12	4	< 0.01
QR	129.64	4	< 0.01
AR	77.57	4	< 0.01
Total	99.52	4	< 0.01
SJT	114.2	4	< 0.01

To understand how age relates to subtest performance, Table 24 shows the correlation between candidate age and their performance on each subtest. The significance column indicates that all the subtests had statistically significant correlations except for VR. For the cognitive subtests with significant correlations, age is slightly negatively correlated with performance, meaning as candidates get older, they tend to perform less well. The strongest correlation is for QR. The correlation is reversed for the SJT. The older a candidate is, the better they tend to perform on the SJT. This makes sense intuitively as candidates who are older might have obtained more of the necessary social skills to exercise appropriate situational judgement.

Table 24. Correlation of Scaled Score with Age (ungrouped)

Subtest	Correlation	Significance
VR	-0.006	$p > 0.01$
DM	-0.089	$p < 0.01$
QR	-0.112	$p < 0.01$
AR	-0.080	$p < 0.01$
Total Cognitive	-0.090	$p < 0.01$
SJT	0.046	$p < 0.01$

Note. Candidates with an age of 14 or below or 56 and above were deemed as invalid and removed from this analysis.

3.5.6 Education

Candidates are requested to state their highest academic qualification, and these are then grouped into the following categories:

1. School leaver qualifications (e.g. A-level, Higher/Advanced Higher, Irish Leaving Cert, IB, BTEC)
2. Degree level or above (e.g. BA, BSc, MA, MSc, PhD)
3. No formal qualifications

The majority of 2021 candidates had a school leaver qualification (80%), nineteen percent had a degree or above, and a small minority had no formal qualifications.

For further analysis, candidates were grouped into those with an honours degree or above and those without an honours degree. Candidates with an honours degree or above performed better on average on VR and the SJT. For the other cognitive subtests and the total cognitive score, below-honours degree candidates performed better on average, as shown in Table 25.

Table 25. Education Scaled Scores

Mean Scaled Score		
Subtest	Below Honours Degree Level (*N* = 30,275)	Honours Degree Level or Above (*N* = 6,942)
VR	570.69	578.42
DM	611.39	604.43
QR	667.93	654.45
AR	652.63	644.99
Total Cognitive	2,502.64	2,482.30
SJT	594.44	611.96
SD		
VR	74.27	78.58
DM	88.32	89.14
QR	80.44	76.12

AR	94.52	93.3
Total Cognitive	275.59	273.06
SJT	74.64	73.1

Table 26 shows that the differences observed in Table 25 are statistically significant.

Table 26. Education *T*-Test

Subtest	*T*-Statistic	*df*	*p* Value
VR	7.73	37,215	< 0.01
DM	-5.91	37,215	< 0.01
QR	-12.71	37,215	< 0.01
AR	-6.08	37,215	< 0.01
Total Cognitive	-5.56	37,215	< 0.01
SJT	17.7	37,215	< 0.01

3.5.7 Country of Residence

Candidates were required to state their country of residence, and these are categorised as UK, EU or Rest of World. The majority of candidates who take the UCAT are resident in the UK, as can be observed in Table 27 below.

Table 27. Candidate Count by Residence

Country of Permanent Residence	*N*	Percent
UK	31,016	83%
Rest of World	4,986	13%
EU	1,215	3%

In past technical reporting, EU and Rest of World are combined into one category called Non-UK. Since 2017 the proportion of candidates who reside in the UK has slightly increased. However, the proportion did not change between 2020 and 2021, as illustrated in Figure 12 below.

Figure 12. Country of Residence 2017–2021

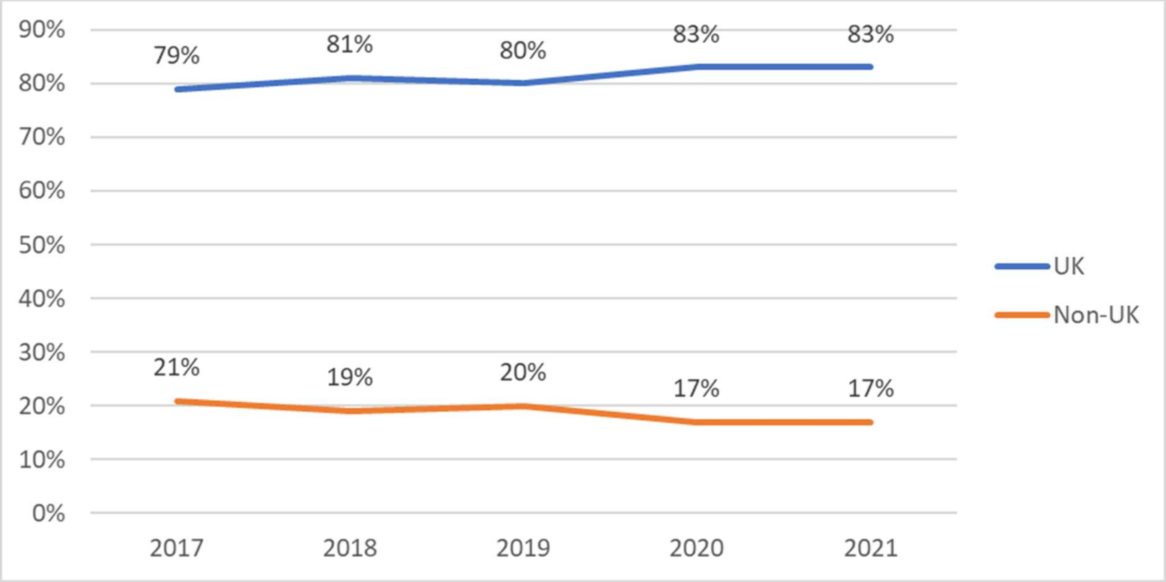


Table 28 shows that UK candidates outperform EU and Rest of World candidates across all subtests.

Table 28. Candidate Scaled Scores by Residence

Mean Scaled Score			
Subtest	UK	Rest of World	EU
VR	575.17	555.39	563.37
DM	613.59	591.52	596.93
QR	667.65	656.96	643.22
AR	655.07	632.06	630.89
Total Cognitive	2,511.48	2,435.94	2,434.41
SJT	604.25	561.77	578.31
SD			
VR	73.99	80.43	72.5
DM	87.03	95.98	83.45
QR	77.75	91.96	72.72
AR	92.85	101.21	90.53
Total Cognitive	268.65	308.08	253.26
SJT	69.72	91.07	76.86

An *F*-test of the differences observed between UK and non-UK candidates is presented in Table 29 below. It shows that the differences are statistically significant.

Table 29. Residence *F*-Test

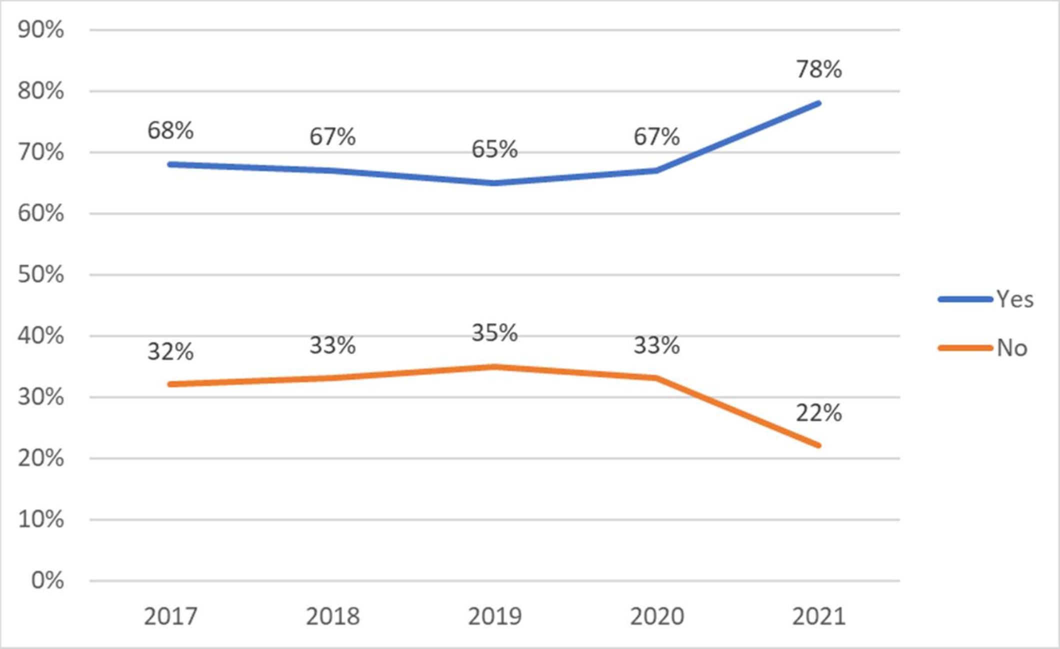
Subtest	*F*-Statistic	*df*	*p* Value
VR	158.67	2	< 0.01
DM	148.53	2	< 0.01
QR	87.42	2	< 0.01
AR	158.24	2	< 0.01
Total Cognitive	198.29	2	< 0.01
SJT	767.76	2	< 0.01

3.5.8 First Language

In 2021 most candidates who sat the UCAT stated that English was their first or primary language (78%). The remaining 22% did not have English as their first or primary language.

Since 2017 the proportion of candidates who state that they speak English as a first or primary language has fluctuated. However, between 2020 and 2021 the proportion increased to of 78%. This is likely to be due to a change in how this question was worded to candidates in 2021.

Figure 13. Count of Language 2017–2021



Across all subtests candidates who stated that English was their first language outperformed those who stated that English was not their first language, as shown in Table 30 below.

Table 30. Language Scaled Scores

Mean Scaled Score		
Subtest	No	Yes
VR	539.85	581.46
DM	574.92	620.25
QR	644.3	671.52
AR	635.34	655.78
Total Cognitive	2,394.42	2,529.01
SJT	571.38	605.31
SD		
VR	71	73.72
DM	87.99	86.03
QR	83.53	77.67
AR	97.73	92.83
Total Cognitive	281.04	266.01
SJT	88.39	68.35

In line with the other demographic categories, a test was carried out to understand whether the differences observed in Table 30 can be considered true reflections of the differences between the two groups. Table 31 shows that that such differences are unlikely to have occurred by chance.

Table 31. Language *T*-Test

Subtest	*T*-Statistic	*df*	*p* Value
VR	45.78	37,215	< 0.01
DM	42.17	37,215	< 0.01
QR	27.71	37,215	< 0.01
AR	17.5	37,215	< 0.01
Total Cognitive	40.18	37,215	< 0.01
SJT	37.23	37,215	< 0.01

3.5.9 Demographic Interactions and SEN

The way demographic characteristics influence UCAT scores is fairly well known. In 2020 Pearson VUE undertook an analysis of variance to explore the interaction between demographic variables and SEN exams. The demographic variables were found to have a significant influence on scores across all cognitive subtests. Furthermore, statistically significant relationships were identified between SEN and qualification on QR and VR, meaning there was an effect of SEN on QR and VR scaled scores, but that effect differs between those that had a high qualification versus a low qualification level. QR scores were also influenced by SEN and SEC together, and SEN and gender together.

The results of these analyses tend to support the statistical testing of each demographic characteristic, that is, that the differences we observe between demographics are true reflections of the differing abilities of the demographic groups. They also tend to show that SEN status does interact with certain demographic characteristics to have a combined influence on scores, although this is only apparent on QR for qualification, SEC and gender; and VR for qualification.

4. Exam Timing Analysis

The section time for each candidate is calculated by summing the item and review time for each item and candidate. Table 32 shows the exam timing for each version of the UCAT.

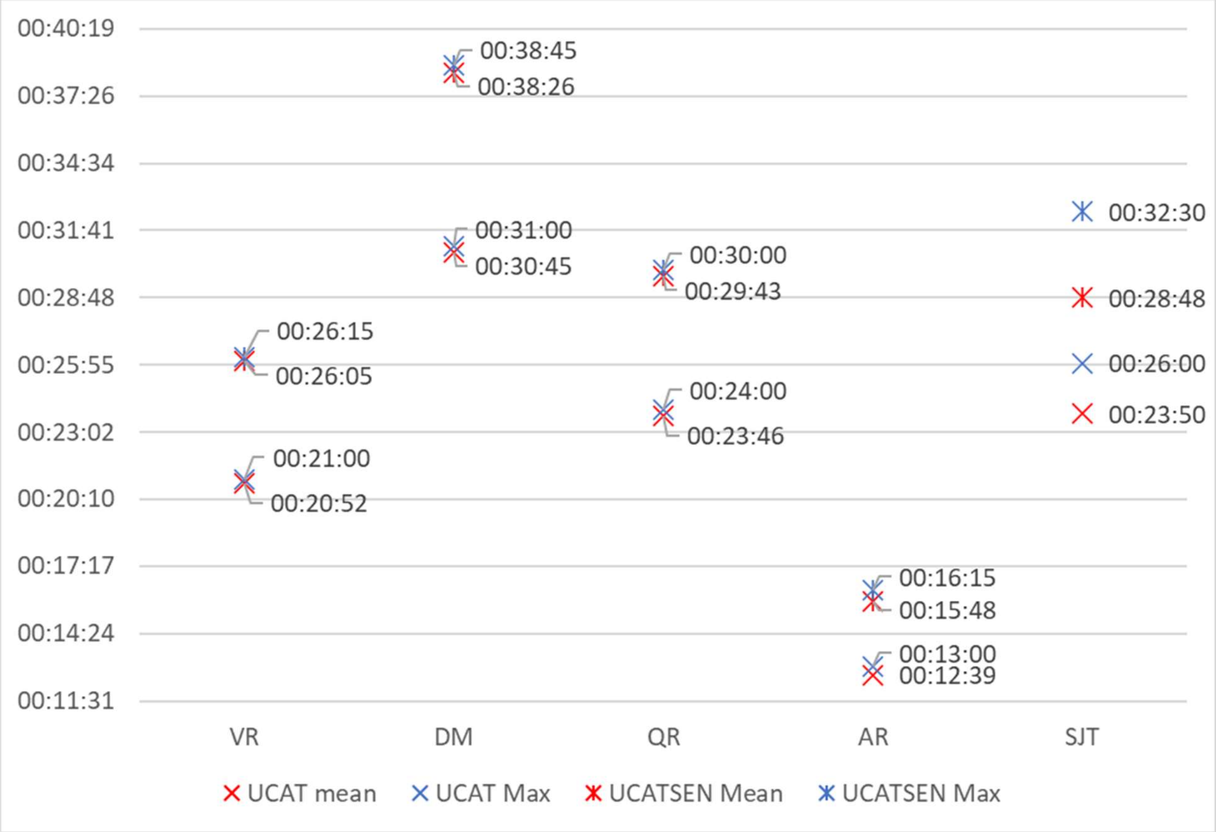
Table 32. Mean Subtest Section Timing: Non-SEN and SEN

Statistic	Subtest	UCAT (32,481)	UCATSEN (1,904)	UCATSENSA (248)	UCATSEN50 (135)	UCATSA (89)
Mean	VR	00:20:52	00:26:05	00:26:04	00:31:04	00:20:53
	DM	00:30:45	00:38:26	00:38:22	00:45:47	00:30:51
	QR	00:23:46	00:29:43	00:29:51	00:35:27	00:23:53
	AR	00:12:39	00:15:48	00:15:55	00:19:00	00:12:40
	SJT	00:23:50	00:28:48	00:28:21	00:32:15	00:24:11
SD	VR	00:00:26	00:00:29	00:00:30	00:02:32	00:00:10
	DM	00:00:56	00:00:59	00:01:14	00:03:39	00:00:18
	QR	00:01:03	00:01:15	00:00:14	00:03:02	00:00:17
	AR	00:00:52	00:01:04	00:00:45	00:01:30	00:00:49
	SJT	00:03:12	00:04:54	00:04:54	00:07:09	00:02:36
Minimum	VR	00:03:41	00:19:56	00:21:41	00:02:38	00:20:13
	DM	00:04:24	00:22:09	00:23:13	00:05:22	00:29:15
	QR	00:00:58	00:02:42	00:28:16	00:01:46	00:21:44
	AR	00:01:38	00:04:37	00:10:34	00:03:26	00:08:46
	SJT	00:01:38	00:04:42	00:11:17	00:12:51	00:13:21
Maximum	VR	00:21:00	00:26:15	00:26:15	00:31:30	00:21:00
	DM	00:31:00	00:38:45	00:38:45	00:46:30	00:31:00
	QR	00:24:00	00:30:00	00:30:00	00:36:00	00:24:00
	AR	00:13:00	00:16:15	00:16:15	00:19:30	00:13:00
	SJT	00:26:00	00:32:30	00:32:30	00:39:00	00:26:00

There is no agreed definition of speededness, although usually it is assessed by examining how closely the average time candidates spend on a subtest is to the total time allowed.

As Table 32 above shows, the cognitive subtests on the UCAT version of the exam are quite speeded. The average time spent completing each subtest is close to the maximum time for each subtest except the SJT, which is considerably less speeded. The SEN versions of the exam are slightly less speeded than the UCAT version. However, the difference between the UCAT version and the UCATSEN version, which is the only SEN version with enough candidates for reliable comparison, is rather small, as show in Figure 14 below. The difference between the average time and the maximum time allowed is barely observable for VR, DM and QR for both UCAT and UCATSEN. The difference is slightly broader for AR and is quite clear for the SJT.

Figure 14. Mean and Maximum Time for UCAT and UCATSEN



Test timing can be examined in more detail in Table 33. It shows that the most speeded non-SEN subtests are VR and QR, where 85% of candidates reached all the items and 7% of candidates did not reach five or more items. The SJT is the least speeded in all exam versions.

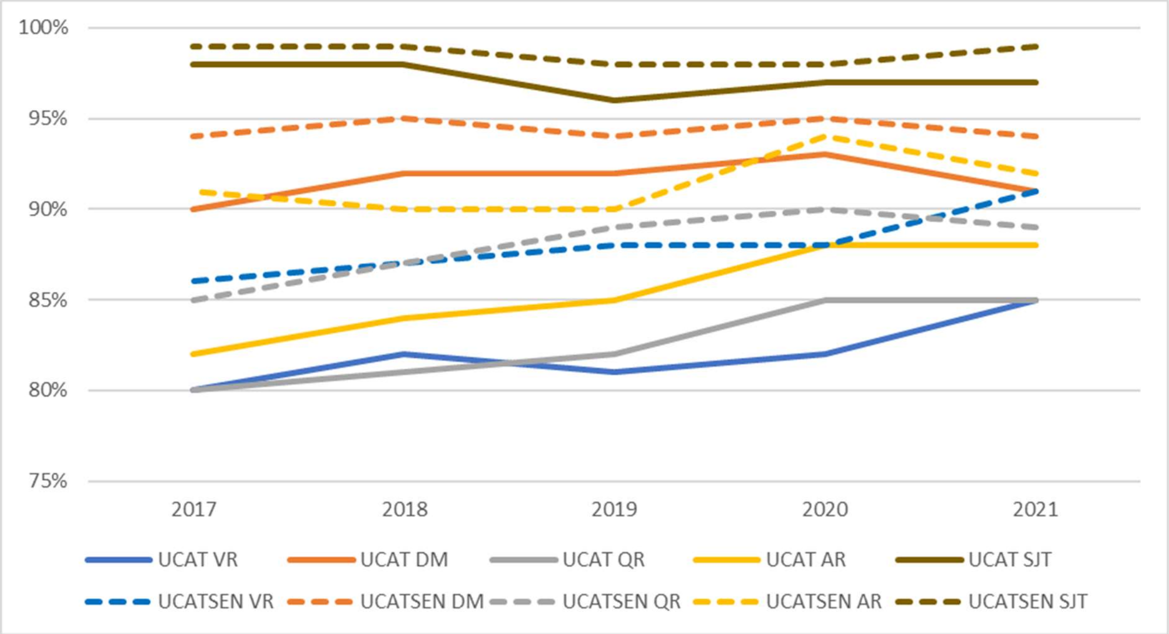
Table 33. Speedeness: Non-SEN and SEN UCAT Candidates

Exam	Subtest	Reached All Items *N*	Reached All Items %	Five or More Items Unreached *N*	Five or More Items Unreached %	Mean Number of Unreached Items for Incomplete Tests Only
UCAT	VR	29,675	85%	2513	7%	6.56 (5166)
	DM	31,732	91%	886	3%	3.7 (3109)
	QR	29,506	85%	2465	7%	6.06 (5335)
	AR	30,545	88%	2267	7%	7.71 (4296)
	SJT	33,776	97%	203	1%	3.7 (1065)
UCATSEN	VR	1,724	91%	72	4%	5.76 (180)
	DM	1,786	94%	26	1%	3.03 (118)
	QR	1,687	89%	91	5%	5.4 (217)
	AR	1,752	92%	72	4%	7.12 (152)
	SJT	1,878	99%	4	0%	2.81 (26)
UCATSENSA	VR	225	91%	9	4%	6 (23)
	DM	238	96%	2	1%	3.7 (10)

Exam	Subtest	Reached All Items *N*	Reached All Items %	Five or More Items Unreached *N*	Five or More Items Unreached %	Mean Number of Unreached Items for Incomplete Tests Only
	QR	220	89%	9	4%	4 (28)
	AR	225	91%	9	4%	6.7 (23)
	SJT	244	98%	0	0%	1.5 (4)
UCATSEN50	VR	125	93%	3	2%	5.1 (10)
	DM	132	98%	0	0%	2.67 (3)
	QR	126	93%	3	2%	3.22 (9)
	AR	130	96%	2	1%	3.4 (5)
	SJT	134	99%	0	0%	1 (1)
UCATSA	VR	83	93%	3	3%	6.67 (6)
	DM	82	92%	1	1%	2.86 (7)
	QR	79	89%	4	4%	4.7 (10)
	AR	83	93%	2	2%	4.67 (6)
	SJT	88	99%	0	0%	1 (1)

Over time, VR, QR and AR have tended to become less speeded, when speededness is defined as the proportion of candidates who reach all the items. Figure 15 shows that although there is a lot of fluctuation year on year, the SJT and DM have fluctuated within a fairly narrow band, whereas the proportion of candidates seeing all the items in the other subtests has gently increased.

Figure 15. Candidates Reaching all Items 2017–2021



5. Test Form Analysis

The 2021 UCAT consisted of five test forms that were delivered randomly to candidates. Table 34 shows the number of candidates who received each form. Forms 1 and 2 received more candidates than the other forms because they were the forms that were delivered to SEN candidates.

Table 34. Candidates by Form

Form	Candidates
Form 1	8,417
Form 2	7,866
Form 3	6,898
Form 4	6,933
Form 5	7,103

Table 35 shows the raw score summary for each subtest on each form. It also includes the reliability statistic, Cronbach's alpha. Alpha is based on the intercorrelations or internal consistency among the items, and it reflects the reproducibility of the test results. High reliability is desirable because it indicates that a test is consistent in measuring the desired construct. AR is consistently the most reliable cognitive subtest, which may be due to the higher number of items. However, all subtests have satisfactorily high reliabilities.

Table 35 also shows *SEM*. This value is the amount of measurement error associated with each subtest and form. *SEM* is calculated using the standard deviation (*SD*) of the raw scores and alpha. Higher reliabilities result in lower *SEM*s.

Table 35. Cognitive Raw Score Test Statistics

Subtest	Form	Mean	*SD*	Min	Max	Alpha	*SEM*
VR (40 items)	Form 1	22.48	5.79	3	39	0.74	2.95
	Form 2	22.93	5.84	2	40	0.73	3.03
	Form 3	21.42	5.92	0	39	0.74	3.02
	Form 4	22.28	5.71	4	39	0.73	2.97
	Form 5	22.32	5.74	2	39	0.73	2.98
DM (26 items)	Form 1	17.23	5.86	1	34	0.77	2.81
	Form 2	16.96	5.86	1	33	0.76	2.87
	Form 3	17.37	5.67	1	32	0.75	2.84
	Form 4	17.54	5.52	1	33	0.73	2.87
	Form 5	16.52	5.16	3	33	0.71	2.78
QR (32 items)	Form 1	18.81	6.21	1	32	0.84	2.48
	Form 2	17.93	5.77	1	32	0.79	2.64
	Form 3	19.55	6.11	1	32	0.83	2.52
	Form 4	18.57	5.66	2	32	0.79	2.59
	Form 5	18.54	5.6	2	32	0.79	2.57

Subtest	Form	Mean	*SD*	Min	Max	Alpha	*SEM*
AR (50 items)	Form 1	32.19	8	1	50	0.84	3.2
	Form 2	31.53	7.94	3	50	0.83	3.27
	Form 3	32.45	8.12	3	50	0.84	3.25
	Form 4	32.35	7.87	2	50	0.84	3.15
	Form 5	30.96	8.26	1	50	0.85	3.2

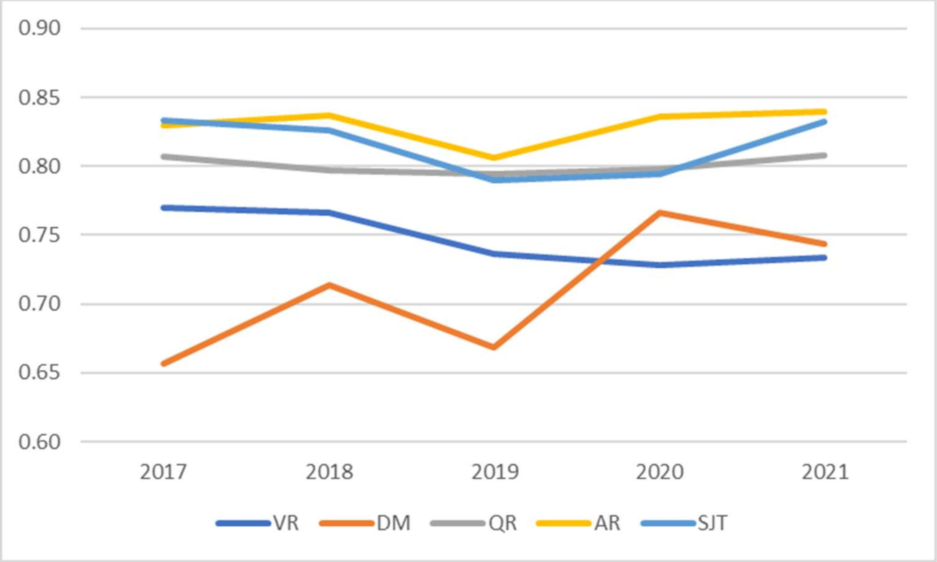
The SJT is analysed in a similar way to the cognitive sections; however, because the maximum raw score available on the SJT can change year on year, an additional column called mean percent raw score is added (Table 36). Similarly to the cognitive results, the reliability is adequately high and the *SEM* adequately low for the SJT.

Table 36. SJT Raw Score Test Statistics

Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Mean Percent Raw Score	Alpha	*SEM*
Form 1	63	8,417	183.42	22.08	42	226	0.75	0.84	8.79
Form 2	63	7,866	180.49	22.67	53	233	0.74	0.85	8.92
Form 3	63	6,898	180.74	20.82	71	223	0.74	0.82	8.88
Form 4	63	6,933	180.47	21.31	50	226	0.74	0.82	9.04
Form 5	63	7,103	178.88	21.45	51	228	0.73	0.83	8.8

Subtest reliability has been fairly consistent since 2017. Figure 16 shows the average Cronbach's alpha for each subtest in each form since 2017. Note that prior to 2019, it is the average of three forms, whereas since 2019, it is the average of five forms. DM has become more reliable since its launch in 2017 and the reliability of VR has slightly dropped, but the other subtests have remained fairly stable.

Figure 16. Raw Score Reliability 2017–2021



Raw scores are scaled and reported as scaled scores. The summary statistics for scaled scores on each form are presented below in Table 37. Instead of alpha, the scaled score reliability is the conditional reliability at each scaled score point. Similarly to the results for raw scores, the scaled score reliability is adequately high for each subtest and each form. Table 37 also includes the results for the SJT.

Table 37. Cognitive Scaled Score Test Statistics

Subtest	Form	Mean	*SD*	Min	Max	Reliability	*SEM*
VR	Form 1	574	75	300	890	0.73	38.83
	Form 2	580	76	300	900	0.74	38.8
	Form 3	561	75	300	890	0.74	38.46
	Form 4	572	73	300	890	0.73	37.97
	Form 5	573	75	300	890	0.73	39.01
DM	Form 1	610	96	300	900	0.79	43.77
	Form 2	608	92	300	890	0.77	43.9
	Form 3	612	87	300	890	0.75	43.41
	Form 4	614	85	300	890	0.74	43.44
	Form 5	607	81	330	890	0.71	43.52
QR	Form 1	668	84	330	900	0.81	36.7
	Form 2	656	76	330	900	0.79	35.01
	Form 3	677	86	330	900	0.81	37.51
	Form 4	664	76	390	900	0.78	35.45
	Form 5	663	74	400	900	0.77	35.57
AR	Form 1	655	94	300	900	0.83	38.76
	Form 2	646	91	300	900	0.82	38.8
	Form 3	658	96	300	900	0.83	39.59
	Form 4	657	93	300	900	0.82	39.63
	Form 5	640	96	300	900	0.84	38.34
Total Cognitive	Form 1	2,507	287	1,380	3,470	0.92	81.12
	Form 2	2,490	272	1,450	3,390	0.91	81.73
	Form 3	2,509	281	1,400	3,420	0.92	79.48
	Form 4	2,507	268	1,450	3,500	0.91	80.35
	Form 5	2,482	265	1,480	3,420	0.91	79.38
SJT	Form 1	608	73	300	749	0.84	28.98
	Form 2	594	76	300	771	0.85	29.93
	Form 3	600	75	300	754	0.82	32.16
	Form 4	594	75	300	755	0.82	31.59
	Form 5	592	73	300	761	0.83	30.11

6. Item Analysis

Each year Pearson VUE undertakes item writing, pretesting, data analysis and statistical screening. New items are pretested along with operational items to establish their efficacy before being introduced into the operational item bank. 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.

The cognitive items are analysed using item response theory whereas the SJT items are analysed using classical test theory, so they are dealt with separately here.

6.1 Cognitive Item Analysis

For the cognitive subtests, quality is assessed on three statistical criteria:

- Point biserial: the degree to which a test item discriminated between strong and weak candidates. For operational items it must be greater than 0.1 for the item to remain in the bank. For pretest items it must be > 0.05.
- *p* Value: the proportion of candidates who answered the item correctly—the item difficulty. This must be between 0.1 and 0.95 for the item to remain in the bank.
- IRT*b*: the difficulty parameter from the item response theory analysis of the items. It must be between -3 and 3 for the item to remain active.

Items that do not meet the statistical criteria laid out above are retired from the bank. It may be possible for them to be revised and reused under a different item ID, but typically they are used for training purposes to show item writers what type of item does not work well.

Table 38 below summarises the number of items that passed the quality criteria by subtest, and by whether they were operational or pretest items. More pretest items tend to fail at this stage since they are new unscored items being tested for the first time. The scored items by contrast have all been previously tested.

Table 38. Cognitive Items Passing the Quality Criteria

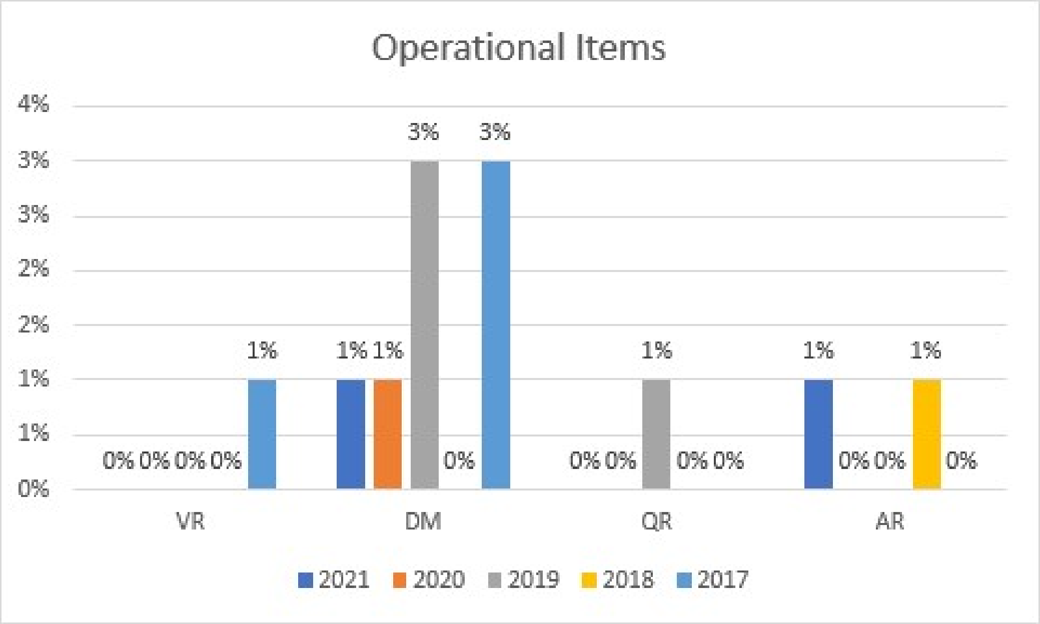
		VR		DM		QR		AR	
		N	%	*N*	%	*N*	%	*N*	%
Operational Scored	Pass	199	100%	129	99%	160	100%	247	99%
	Fail	1	0%	1	1%	0	0%	3	1%
	p < 10 or > 95	0	0%	0	0%	0	0%	0	0%
	pBis <= 0.1	1	0%	1	1%	0	0%	3	1%
	|b| >= 3	0	0%	0	0%	0	0%	0	0%
Pretest Unscored	Pass	228	98%	258	98%	210	96%	232	80%
	Fail	4	2%	6	2%	8	4%	57	20%
	p < 10 or > 95	0	0%	1	0%	5	2%	2	1%

		VR		DM		QR		AR	
		N	%	*N*	%	*N*	%	*N*	%
	pBis <= 0.05	4	2%	5	2%	4	2%	56	19%
	|b| >= 3	0	0%	0	0%	4	2%	8	3%

Consistent with previous years, only five operational items failed the analysis. Those items did not discriminate highly enough. For the pretest items few failed in the VR, DM and QR subtest. On VR pretest failure was due solely to low item discrimination. For DM and QR it was due to low discrimination in addition to items being too easy or difficult. Rather more pretest items failed the AR analysis. The majority failed due to poor discrimination, but there were also some that were not an appropriate difficulty.

The higher rate of failures in the AR subtest is consistent with previous years. As Figure 17 shows, the failure rate for AR pretest items has ranged from 13% to 22% since 2017. Therefore, it appears that it is particularly difficult to produce AR items that work well. However, the low AR operational failure rate demonstrates that pretesting is working effectively at removing poor quality items before they enter the scored item bank.

Figure 17. Proportion of Items Failing Analysis 2017–2021



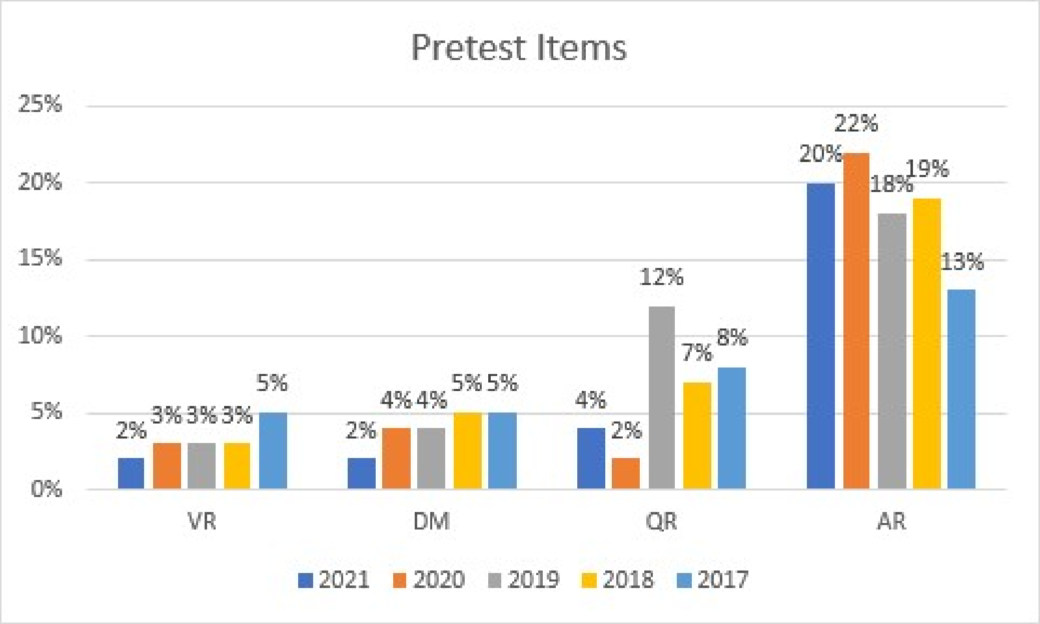


Table 39 shows a summary of the point biserial values. The maximum point biserial is 1, and higher values are better because they indicate that an item can discriminate well between strong and weak candidates. Given that the unscored items have not been tested before, it is expected that those items, on average, will discriminate less well than the scored items, and that is the case across all the cognitive subtests.

Table 39. Discrimination Summary Statistics

Scored/Unscored	Subtest	*N* Items	Mean pBis	*SD* pBis	Min pBis	Max pBis
Operational (Scored)	VR	200	0.28	0.06	0.06	0.45
	DM	130	0.36	0.09	0.04	0.68
	QR	160	0.36	0.06	0.15	0.5
	AR	250	0.33	0.07	0.03	0.48
Pretest (Unscored)	VR	232	0.25	0.08	0.01	0.42
	DM	264	0.3	0.12	0.01	0.64
	QR	218	0.27	0.1	-0.01	0.48
	AR	289	0.15	0.1	-0.19	0.4

Historically the point biserial values for scored items have been high and stable, whereas the values for unscored items have been lower and less consistent, as illustrated in Figure 18. The operational items appear to have become slightly more discriminating over time for all subtests except VR. This is an indication that the quality of the subtest has improved over time.

Figure 18. Point biserial 2017–2021

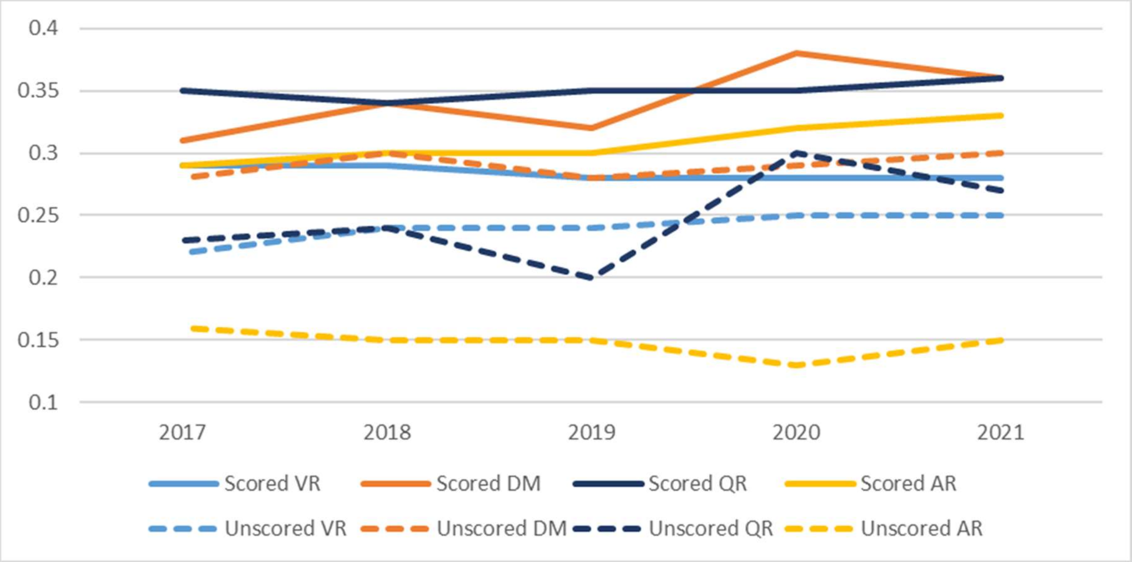


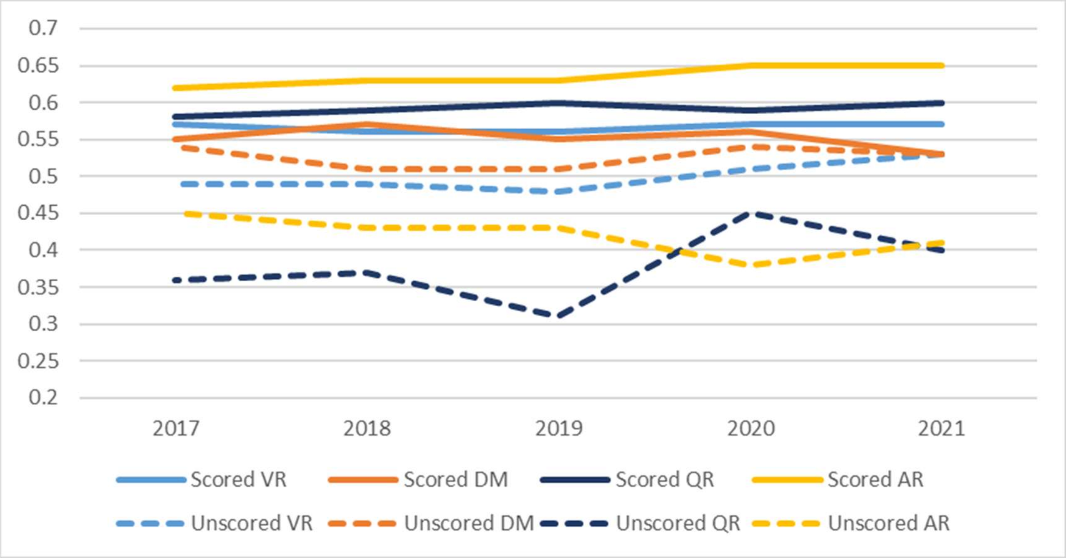
Table 40 shows the summary *of p* values for the cognitive subtests. *p* values reflect the proportion of candidates who answered an item correctly, so higher values indicate easier items, and lower values more difficult items. Of the operational items, DM items appear to have been the most difficult on average for 2021 candidates and AR items were the easiest on average. The pretest pools appear to have been somewhat more difficult overall than the operational test items for all subtests except DM, where they were the same difficulty on average.

Table 40. *p* Value Summary Statistics

Scored/Unscored	Subtest	*N* Items	Mean *p*	*SD p*	Min *p*	Max *p*
Operational (Scored)	VR	200	0.57	0.14	0.25	0.84
	DM	130	0.53	0.17	0.15	0.87
	QR	160	0.6	0.13	0.2	0.86
	AR	250	0.65	0.14	0.21	0.9
Pretest (Unscored)	VR	232	0.53	0.16	0.15	0.9
	DM	264	0.53	0.2	0.08	0.93
	QR	218	0.4	0.17	0.03	0.89
	AR	289	0.41	0.17	0.08	0.91

Since 2017, pretesting has been successful in identifying items that are too difficult and too easy. Figure 19 shows that the items in the pretest pools are usually more difficult than the operational items on average. Note that the subtests are equated year-on-year, meaning changes in difficulty of individual items does not have an impact on the ability required for candidates to achieve a given scaled score.

Figure 19. *p* Value 2017–2021



The VR subtest consists of four-option multiple-choice items and three-option true/false/can't tell items. Table 41 shows that the four-option multiple-choice items are better at discriminating between stronger and weaker candidates than the three-option items. The lower point biserials in the pretest pool shows that pretesting is successfully removing items that do not discriminate effectively. The operational items are also rather easier on average than the pretest pool items.

Table 41. VR Type Point biserial and *p* Value

Scored/Unscored	Item Type	*N* Items	Mean pBis	*SD* pBis	Mean *p*	*SD p*
Operational (Scored)	Multiple Choice	120	0.31	0.05	0.57	0.12
	True/False/Can't Tell	80	0.24	0.05	0.57	0.16
Pretest (Unscored)	Multiple Choice	192	0.26	0.08	0.53	0.16
	True/False/Can't Tell	40	0.22	0.07	0.54	0.17

The DM subtest contains multiple-choice items, scored out of 1 and drag-and-drop items, which are scored out of 2. The drag-and-drop items are more difficult than the multiple-choice items although they discriminate less well, as shown in Table 42. Coincidentally the average point biserial for operational drag-and-drop items is the same as the *p* value for operational drag-and-drop items.

Table 42. DM Type Point biserial and *p* Value

Scored/Unscored	Item Type	*N* Items	Mean pBis	*SD* pBis	Mean *p*	*SD p*
Operational (Scored)	Drag and Drop	40	0.44	0.09	0.44	0.14
	Multiple Choice	90	0.32	0.07	0.57	0.17
Pretest (Unscored)	Drag and Drop	67	0.38	0.12	0.49	0.19
	Multiple Choice	197	0.27	0.1	0.54	0.2

The QR subtest has item sets and standalone items. Each item set contains four items. As with the pretest pool as a whole, the untested items discriminate less well on average than the ones that have already been pretested prior to appearing in the 2021 exam, as shown in Table 43.

Table 43. QR Type Point biserial and *p* Value

Scored/Unscored	Item Type	*N* Items	Mean pBis	*SD* pBis	Mean *p*	*SD p*
Operational (Scored)	Item Set	140	0.37	0.06	0.59	0.13
	Standalone	20	0.34	0.05	0.65	0.15
Pretest (Unscored)	Item Set	198	0.27	0.1	0.39	0.16
	Standalone	20	0.27	0.09	0.52	0.19

The AR subtest consists of four different types. Table 44 below shows that the discrimination of the operational items are higher than the pretest items, which demonstrates that pretesting is working at removing the items that do not work very well.

Table 44. AR Type Point biserial and *p* Value

Scored/Unscored	Item Type	*N* Items	Mean pBis	*SD* pBis	Mean *p*	*SD p*
Operational (Scored)	Type 1	200	0.33	0.07	0.66	0.14
	Type 2	10	0.27	0.08	0.62	0.19
	Type 3	15	0.28	0.06	0.6	0.17
	Type 4	25	0.37	0.05	0.59	0.1
Pretest (Unscored)	Type 1	150	0.16	0.11	0.47	0.17
	Type 2	19	0.15	0.1	0.39	0.16
	Type 3	20	0.19	0.09	0.46	0.2
	Type 4	100	0.11	0.07	0.31	0.08

6.2 SJT Item Analysis

Unlike the analysis undertaken on the cognitive sections, 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.

Prior to calculating the item statistics, outlier candidates are removed from the sample according to the criteria outlined in Table 45. The candidates that are removed are judged as not interacting with the test as expected and are therefore not representative of the UCAT population.

Table 45. Candidate Removal Summary for SJT Item Analysis

Statistic	Criteria	Number of Candidates Removed
1. *Z* score of the scaled score	*Z* score < -4.17894	0
2. High number of missing responses	> 1 blank response on operational items	1,259
3. Low completion time	Drop in score based on response time	0

The following item statistics are calculated for the SJT items:

- Item facility: the mean score on the items as a percentage of the maximum score available. It represents the difficulty of the item.
- Item *SD*: the *SD* of the scores on the items. It gives an indication of how well the item is differentiating among candidates.
- Item partial correlation: the correlation of the item score with the total score for the operational items and the scaled score for the pretest items. It compares how individuals perform on a given item with how they perform on the test overall and

is a measure of discrimination. Item correlations can be interpreted in the following way:

- Below 0.13 – poor correlation with the test overall and items within this band are unlikely to be used in an operational test.
- 0.13 to 0.17 – acceptable correlations. Items within this band will only be included if other items within the scenario have higher item partials.
- 0.17 to 0.25 – reasonable item performance.
- Above 0.25 – good item performance.

SJT items should meet the following quality criteria:

- Item facility < 95%
- Item *SD* >= 0.30
- Item partial >= 0.13

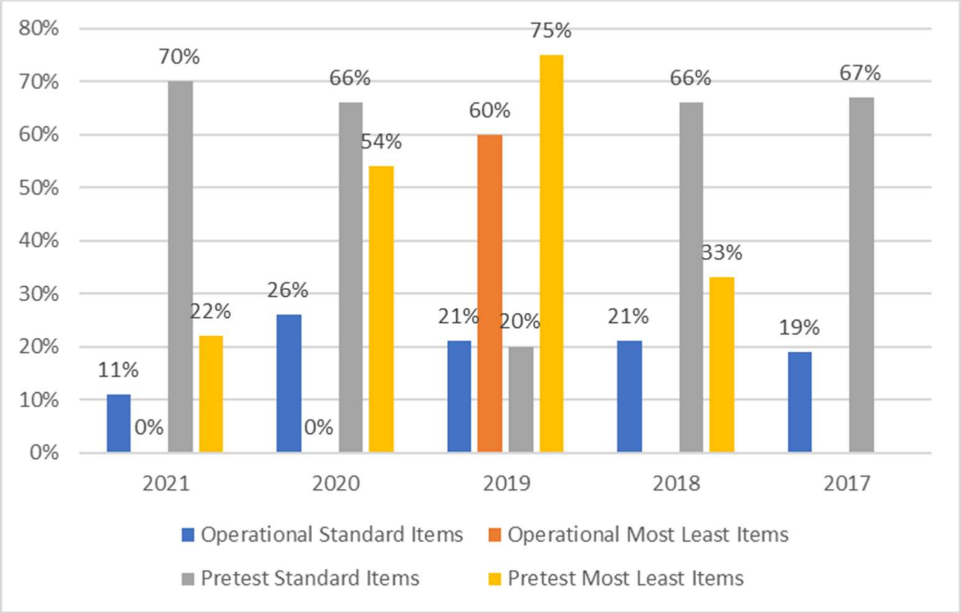
Every effort is made not to use items that do not meet these criteria on future forms, although items with high partials and borderline facility and/or *SD* values are sometimes retained. Table 46 shows the number of items that met and did not meet the quality criteria. The most/least item type was more successful than the standard items, with all operational items and 78% of the pretest items meeting the criteria.

Table 46. SJT Items Quality Criteria

	Item Type	Statistical Criteria Met/Not Met	All		Appropriateness		Importance		Direct Speech	
			N	%	*N*	%	*N*	%	*N*	%
Operational	Standard Items	Met	159	89%	59	100%	75	86%	25	76%
		Not met	20	11%	19	0%	12	14%	8	24%
	Most/Least Items	Met	5	100%						
		Not met	0	0%						
Pretest	Standard Items	Met	73	30%	67	30%	1	17%	5	31%
		Not met	170	70%	154	70%	5	83%	11	69%
	Most/Least Items	Met	7	78%						
		Not met	2	22%						

The proportion of items meeting the quality criteria is fairly consistent with previous years. Figure 20 illustrates that the proportion of operational standard items not meeting the criteria in 2021 fell to 11% and the number of pretest most/least items not meeting the criteria fell to 22%; however, the number of standard pretest items not meeting the criteria increased to 70%.

Figure 20. Proportion of Items Failing Analysis 2017–2021



The summary of all operational SJT items is shows below in Table 47.

Table 47. Operational SJT Item Analysis Summary

Items: 203	Mean	*SD*	Min	Max
Item mean	2.83	1.04	0.83	7.32
Item *SD*	1.05	0.31	0.3	1.8
Item partial correlation	0.25	0.12	-0.15	0.51
Item total facility	72.93	16.94	27.57	98.79

Since 2017 the item mean score and facility has tended to increase, as illustrated in Figure 21 below, indicating that items are becoming somewhat easier. The total number of items has also progressively increased since 2017. The increase in item partial correlation shows that the items are getting better overall at discriminating among strong and weak candidates.

Figure 21. SJT Operational Summary 2017–2021

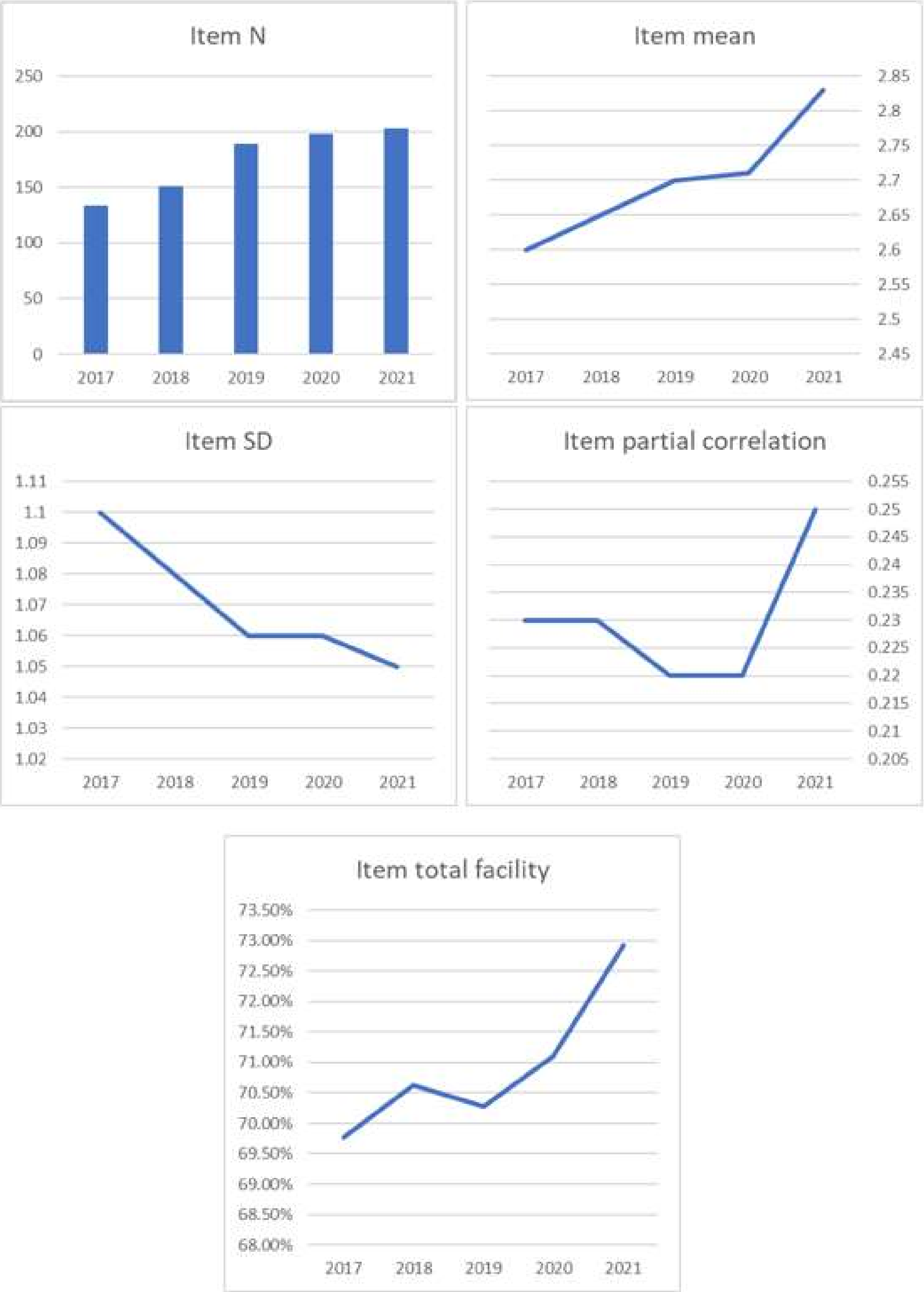


Table 48 shows the summary statistics for the SJT pretest items.

Table 48. SJT Pretest Item Summary Statistics

	Statistic	Item Mean	Item *SD*	Item Mean	Item Total Facility
Rating Items (243 items)	Mean	3.01	0.85	0.09	79.58%
	SD	0.83	0.3	0.14	15.90%
	Min	0.25	0.09	-0.23	8.18%
	Max	3.99	1.55	0.42	99.80%
Most/Least (9 items)	Mean	7.23	1.3	0.16	90.42%
	SD	0.48	0.43	0.04	6.02%
	Min	6.23	0.87	0.1	77.82%
	Max	7.62	2.23	0.23	95.23%

6.3 Differential Item Functioning

6.3.1 Introduction

DIF is a method for detecting potential bias in test items. 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 characteristic of the candidates that is related to gender.

The UCAT DIF comparison groups are based on gender, age, ethnicity, SEC, level of education, first language, permanent residence, and mode of delivery.

6.3.2 Method of DIF Detection

For the 2021 UCAT a different method of DIF detection was employed for the cognitive sections and the SJT due to the different measurement models employed by the subtests. For the cognitive subtests the Mantel-Haenszel procedure was used. This procedure compares the performance of different groups of candidates who are within the same ability strata. If there are overall differences between the groups for candidates of the same ability levels, then the item may be measuring something other than what it was designed to measure.

Since the SJT makes extensive use of polytomous scoring, the DIF analysis was performed with a hierarchical regression approach using the equated scaled score.

In both approaches, 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: DIF is not significantly different from zero or has an absolute value < 1.0

B: DIF is significantly different from zero and has an absolute value >= 1.0 and < 1.5
C: DIF is significantly larger than 1.0 and has an absolute value >= 1.5

Items flagged in Category C are removed from the item bank on the basis that they may contain bias. Items flagged in Categories A and B are not removed because of the small effect or lack of statistical significance.

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.

6.3.3 Sample Size Requirements

Minimum sample-size requirements used for the UCAT DIF analyses were at least 50 candidate responses per group and at least 200 in total. If the sample size for the DIF analysis is less than 200, the sample is not large enough to undertake analysis and therefore DIF is not reported. 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 certain group comparisons.

6.3.4 DIF Results

The DIF results are now reported for each demographic group. Table 49 shows DIF in relation to gender. Two operational items were found to exhibit Category C DIF in the DM subtest. These were both items that favoured males over females. One pretest item was also found that exhibited Category C DIF. It favoured females over males. The items were reviewed by the content development team to identify whether bias is likely to be the source of the DIF and removed from the item bank so they cannot be used in future iterations of the test.

Table 49. Gender DIF

Group	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	197	98%	124	95%	159	99%	248	99%	199	98%
	B	3	2%	4	3%	1	1%	2	1%	4	2%
	C	0	0%	2	2%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	A	232	100%	263	100%	217	100%	289	100%	242	96%
	B	0	0%	0	0%	1	0%	0	0%	10	4%
	C	0	0%	1	0%	0	0%	0	0%	0	0%

	NA	0	0%	0	0%	0	0%	0	0%	0	0%

In the age comparison, DIF could not be reliably calculated for many of the items. This was due to the low number of candidates in the over 35 age group. Only 301 candidates were aged over 35 (as discussed in 3.5.5 above), which explains why many items are categorised as NA. One Category C item was identified in VR. It favoured under 20-year-olds over people older than 35.

Table 50. Age DIF

Group	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	74	37%	46	35%	57	36%	85	34%	198	98%
	B	0	0%	0	0%	0	0%	0	0%	5	2%
	C	1	0%	0	0%	0	0%	0	0%	0	0%
	NA	125	62%	84	65%	103	64%	165	66%	0	0%
Pretest	A	0	0%	0	0%	0	0%	0	0%	246	98%
	B	0	0%	0	0%	0	0%	0	0%	6	2%
	C	0	0%	0	0%	0	0%	0	0%	0	0%
	NA	232	100%	264	100%	218	100%	289	100%	0	0%

For ethnicity, there were usually enough items to reliably categorise DIF for operational items. However, because there were more pretest items, many of the pretest comparisons are not possible due to low candidate numbers. For instance, there were only 469 UK – Chinese candidates.

Table 51 shows there were five instances of C DIF identified in the ethnicity comparisons. Two were operational QR items, which were found to favour White over Black in one instance and Black over White in the other. One Category C item was found in the operational White/Chinese comparison. It was in the DM subtest and favoured Chinese over White candidates. Two pretest items was found to exhibit Category C DIF. One was in QR subtest and the other in the SJT; they both favoured White over Asian candidates.

Table 51. Ethnicity DIF

Type	Group	Code	VR		DM		QR		AR		SJT	
			N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	White/ Black	A	191	96%	126	97%	155	97%	248	99%	189	93%
		B	9	4%	4	3%	3	2%	2	1%	14	7%
		C	0	0%	0	0%	2	1%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
	White/ Asian	A	195	98%	127	98%	160	100%	249	100%	182	90%
		B	5	2%	3	2%	0	0%	1	0%	21	10%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
	White/ Chinese	A	199	100%	128	98%	160	100%	250	100%	200	99%
		B	1	0%	1	1%	0	0%	0	0%	3	1%
		C	0	0%	1	1%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
	White/ Mixed	A	199	100%	130	100%	160	100%	250	100%	203	100%
		B	1	0%	0	0%	0	0%	0	0%	0	0%

		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	White/ Black	A	185	80%	38	14%	192	88%	221	76%	39	15%
		B	0	0%	0	0%	0	0%	0	0%	2	1%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	47	20%	226	86%	26	12%	68	24%	211	84%
	White/ Asian	A	232	100%	250	95%	217	100%	289	100%	240	95%
		B	0	0%	0	0%	0	0%	0	0%	10	4%
		C	0	0%	0	0%	1	0%	0	0%	1	0%
		NA	0	0%	14	5%	0	0%	0	0%	1	0%
	White/ Chinese	A	0	0%	0	0%	0	0%	0	0%	9	4%
		B	0	0%	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	232	100%	264	100%	218	100%	289	100%	243	96%
	White/ Mixed	A	0	0%	0	0%	0	0%	1	0%	9	4%
		B	0	0%	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	232	100%	264	100%	218	100%	288	100%	243	96%

No Category C DIF was identified in the SEC comparisons for the operational items. For operational items, there were plenty of candidate responses to reliably categorise items appropriately, but as Table 52 demonstrates, very few comparisons were possible for the pretest items. Nonetheless, three pretest items were found to exhibit Category C DIF. In the SJT, two items favoured SEC 1 over SEC 2, and one favoured SEC 1 over SEC 5.

Table 52. SEC DIF

Type	Group	Code	VR		DM		QR		AR		SJT	
			N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	SEC 1/2	A	200	100%	130	100%	160	100%	250	100%	203	100%
		B	0	0%	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
	SEC 1/3	A	200	100%	130	100%	160	100%	250	100%	203	100%
		B	0	0%	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
	SEC1/4	A	200	100%	130	100%	160	100%	250	100%	203	100%
		B	0	0%	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
	SEC 1/5	A	199	100%	130	100%	160	100%	250	100%	203	100%
		B	1	0%	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	SEC 1/2	A	0	0%	0	0%	0	0%	0	0%	56	22%
		B	0	0%	0	0%	0	0%	0	0%	1	0%
		C	0	0%	0	0%	0	0%	0	0%	2	1%
		NA	232	100%	264	100%	218	100%	289	100%	193	77%
	SEC 1/3	A	0	0%	0	0%	2	1%	2	1%	0	0%
		B	0	0%	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%	0	0%

		NA	232	100%	264	100%	216	99%	287	99%	252	100%
	SEC 1/4	A	0	0%	0	0%	0	0%	0	0%	78	31%
		B	0	0%	0	0%	0	0%	0	0%	2	1%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	232	100%	264	100%	218	100%	289	100%	172	68%
	SEC 1/5	A	0	0%	0	0%	2	1%	1	0%	121	48%
		B	0	0%	0	0%	0	0%	0	0%	1	0%
		C	0	0%	0	0%	0	0%	0	0%	1	0%
		NA	232	100%	264	100%	216	99%	288	100%	129	51%

As Table 53 illustrates, there was no Category C DIF detected in the comparison between candidates who had an honours degrees or above and those who did not. There were high candidate volumes across the board, meaning comparisons could be made for all subtests.

Table 53. Honours Degree DIF

Type	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	199	100%	127	98%	159	99%	250	100%	200	99%
	B	1	0%	3	2%	1	1%	0	0%	3	1%
	C	0	0%	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	A	232	100%	264	100%	218	100%	289	100%	245	97%
	B	0	0%	0	0%	0	0%	0	0%	7	3%
	C	0	0%	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%	0	0%

Comparison was also possible for the most part across all subtests for candidates who reported English as being their first or primary language and those who reported that it was not. As Table 54 shows, only one item could not be categorised—a DM pretest item. No Category C DIF was detected.

Table 54. English as First Language DIF

Group	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	200	100%	130	100%	160	100%	250	100%	202	100%
	B	0	0%	0	0%	0	0%	0	0%	1	0%
	C	0	0%	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	A	232	100%	263	100%	218	100%	289	100%	239	95%
	B	0	0%	0	0%	0	0%	0	0%	13	5%
	C	0	0%	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	1	0%	0	0%	0	0%	0	0%

One Category C DIF item was identified in the comparison of candidates who reported UK as their residence with those who reported the UK as not being their residence. It was a QR pretest item that favoured UK residence candidates over non-UK residence candidates.

Table 55. Residency DIF

Group	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	200	100%	129	99%	159	99%	249	100%	203	100%
	B	0	0%	1	1%	1	1%	1	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	A	232	100%	229	87%	217	100%	289	100%	236	94%
	B	0	0%	0	0%	0	0%	0	0%	16	6%
	C	0	0%	0	0%	1	0%	0	0%	0	0%
	NA	0	0%	35	13%	0	0%	0	0%	0	0%

Very few candidates took the online version of the UCAT (231 candidates, see 3.4 above), so comparison was not possible on many operational items, and no pretest items could be reliably analysed. As Table 56 shows, no Category C DIF was detected.

Table 56. Delivery Mode DIF

Group	Code	VR		DM		QR		AR	
		N	%	*N*	%	*N*	%	*N*	%
Operational	A	77	38%	50	38%	63	39%	100	40%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	123	62%	80	62%	97	61%	150	60%
Pretest	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	232	100%	264	100%	218	100%	289	100%

In summary, 13 items were found to exhibit Category C DIF. These items were removed from the bank so they will not be selected for future forms.

7. Recommendations

The outcome of the UCAT 2021 analysis identifies certain operational changes that could improve the ongoing performance of the test, as well as several areas that might provide fertile ground for further research.

As it stands certain subtests have a greater impact on the total cognitive score that candidates receive than others. Specifically, QR, as the highest scoring subtest, has a greater influence on the total score than VR, which is the lowest scoring subtest. Pearson VUE proposed to slightly rescale the higher scoring subtests year-on-year to bring them closer to an average score of 600. This activity should continue with both QR and AR being rescaled to reduce their disproportionate impact on the total cognitive scaled score.

The speededness of the cognitive subtests has fallen slightly on some subtests, but the degree of speededness is still potentially a concern. Currently Pearson VUE constructs test forms with a constraint on the historic average response time of each item for AR and QR. This constraint means items that take a long time to answer are not included in the forms. As a result, the average time to answer all items on the subtest is kept to a reasonable level. Although the DM and VR subtests do not currently show excessive speededness, Pearson VUE will apply a time constraint on form construction for those subtests in future, as a good way to pre-emptively control the degree of speededness.

Item time also presents an opportunity to examine the influence of time available on performance. For many aptitude tests, speededness is part of the test construct, in the sense that it encourages candidates to rely on their natural aptitude and not answer the questions by applying knowledge or test wisdom strategies. A potentially fruitful area of research would be to examine the degree to which increased time influences scores, all else being equal, and furthermore, what cognitive strategies candidates use under different time constraints. The outcome of such research would help inform discussions on the time allowed for each section, and potentially allow for well-performing but time-consuming items to be reintroduced into the operational exam.

Examination of the relative overperformance of SEN candidates also relates to time allowed. In 3.2 above it was stated that some of the difference in scores was found to be related to demographic characteristics of candidates who take SEN versions of the exam. However, score differences remained even after these demographic differences were accounted for. Further research could improve our knowledge on the appropriate time allowed by examining who the SEN candidates are, and how they take the test.

One way to understand better who the SEN candidates are is to better understand subgroup differences in scores. There are several characteristics that are collected but not reported here, such as whether candidates receive free school meals or a bursary. Pearson VUE will examine these characteristics and continue to explore ways to better understand subgroup differences.

Finally, non-UK candidates tend to perform less well on the SJT than UK candidates. The hypothesis put forward in 3.5.3 above was that this may be due to situational judgement being linked to geographical cultural competence. However, the absence of category C DIF (see 6.3.4) would tend to indicate that whatever separates these two groups is integral to the measurement of situational judgement. In aiming to reduce the difference between UK and non-UK candidates on the SJT it would be useful to examine the degree to which it is possible to separate content relevant to good situational judgement from content that may not be immediately accessible to non-UK candidates.