Pearson

University Clinical Aptitude Test (UCAT)

Technical Report
Testing Interval: 10 July 2023 to 28 September 2023

Prepared by:
Pearson VUE
19 March 2024

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

1. Executive Summary

The University Clinical Aptitude Test (UCAT) was administered in 2023 from 10 July 2023 to 28 September 2023. This report covers the 35,625 exams that were delivered during that period, which is a small decrease (2%) from 2022. The exam was delivered in two modes: online and test centre. Online test delivery accounted for only 0.1% of candidates, so it is not possible to reliably compare results between these two groups.

This report covers four of the five versions of the UCAT made available for candidates with special educational needs (SEN). One version, taken during the contingency period, is not included in this report. 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. The scaling of the subtests in 2023 was adjusted to even out the distribution of scaled scores among these subtests. This adjustment resulted in a higher mean scaled score for Verbal Reasoning (VR) and lower scores for Quantitative Reasoning (QR) and Abstract Reasoning (AR), bringing the averages of all subtests closer together. After accounting for the rescaling effort, the mean scaled scores for VR, QR, Decision Making (DM), and AR remained stable and were comparable to the mean scaled scores in 2022. The Situational Judgement Test (SJT) bands showed a deviation within 4% of the target proportions. Notably, the percentage of candidates in the lowest SJT band fell from 14% in 2022 to 9% in 2023, aligning more closely with the target of 10%.

The 2023 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 speeded to a certain extent. Most candidates used all the available time and the average time used was very close to the available time. In 2022, changes were made to lessen the time pressure in these subtests, and these changes continued into 2023. As a result, the level of time pressure in 2023 was similar to 2022 but lower than in previous years. Analysis excluding guesses suggests that candidates were generally able to attempt most questions in each subtest. Speededness was lower in the SEN exams, where candidates have more time available. The SJT remains the least speeded subtest.

In 2023, demographic trends largely mirrored those of past years, with the notable exceptions of a continuous decrease in UK - White candidates and a rise in UK - Asian and non-UK candidates. Candidates with a higher socio-economic classification (SEC), those of white ethnicity, English as a first language speakers, and UK residents were associated with higher scores. In the cognitive subtests, male candidates generally outperformed female candidates, while in the SJT, female candidates performed better than their male counterparts.

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. Four operational items and 17 pretest items from the cognitive subtests did not meet quality standards and were removed from the item bank. In the SJT subtest, 35 operational items and 195 pretest items were found to have failed to meet all of the relevant criteria. Additionally, 8 operational items and 16 pretest items were removed due to potentially exhibiting bias.

2. 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 2023 UCAT that was delivered from 10 July 2023 to 28 September 2023. As outlined in Section 3, the exam consisted of five subtests ranging from 29 to 69 items each. The design of the exam remained the same as in the previous year, with a small change to the scaling of three of the subtests. The VR subtest was scaled up by 20 scaled score points while the QR subtest and AR subtest were scaled down by 10 scaled score points each.

Section 4 describes the exam results in terms of candidate volumes, scaled scores, and SJT bands. It also 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 5. Section 6 contains the analysis of the five test forms, Section 7 summarises the analysis of the test items, and the final section of this report provides recommendations for future testing cycles.

3. Exam Design 2023

The 2023 UCAT consisted of five balanced test forms, each with five subtests. Each subtest includes scored and unscored items as shown in Table 1 below, in addition to changes made to the 2023 test.

Table 1. 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
DM	1 testlet of 26 items	3 items	29	31 minutes allowed on items and 1 minute for instruction
QR	8 testlets of 4 items	1 testlet of 4 items	36	25 minutes allowed on items and 1 minute for instruction
AR	10 testlets of 5 items	0 items	50	12 minutes allowed on items and 1 minute for instruction
SJT	20 testlets of 1 to 4 items	2 testlets of 1 to 5 items	69	26 minutes allowed on items and 1 minute for instruction

Candidates were given 120 minutes to answer a total of 228 items from the five subtests. There were five groups of candidates who took a SEN version of the exam, and thus had extra time allowances in 2023. The timing and scoring of the SEN exams are explored in detail in Section 4.2.

There have been changes to the scaling of the subtests in 2023. For the past 5 years, the mean scaled scores for QR and AR were comparatively higher than the other subtests, while for VR, the mean scaled score was relatively lower. Therefore, UCAT decided to scale down both QR and AR by 10 points and scale up VR by 20 points to narrow the gap between the cognitive subtests while maintaining similar total cognitive subtest scores.

The raw scores in each cognitive subtest were transformed to a scaled score ranging from 300 to 900. SJT scaled scores ranged from 300 to 790. Universities received the 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.

Table 2. SJT Band Scaled Score Range and Description

Band	Scaled Score Range	Intended Band Proportions	Narrative
Band 1	656–900	22%	Those in Band 1 demonstrated an excellent level of performance, showing similar judgement in most cases to the panel of experts.
Band 2	593–655	38%	Those in Band 2 demonstrated a good, solid level of performance, showing appropriate judgement frequently, with many responses matching model answers.
Band 3	495–592	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	300–494	10%	The performance of those in Band 4 was low, with judgement tending to differ substantially from ideal responses in many cases.

The 2023 UCAT was delivered in two modes: the OnVUE mode, where a candidate can take the test remotely with an online proctor, or the test centre mode, where candidates take the test in a specially designed test centre. Only 31 candidates took the online version of the test (see Section 4.4).

4. Examination Results

4.1 Overall Exam Results

This report covers examination results for the 35,625 candidates who took the UCAT during the period 10 July 2023 to 28 September 2023. Candidate volumes have increased each year from 2017 to 2021 but showed a small decrease of 4% from 2021 to 2023, as illustrated in Figure 1 below.

Figure 1. Candidate Volumes since 2017

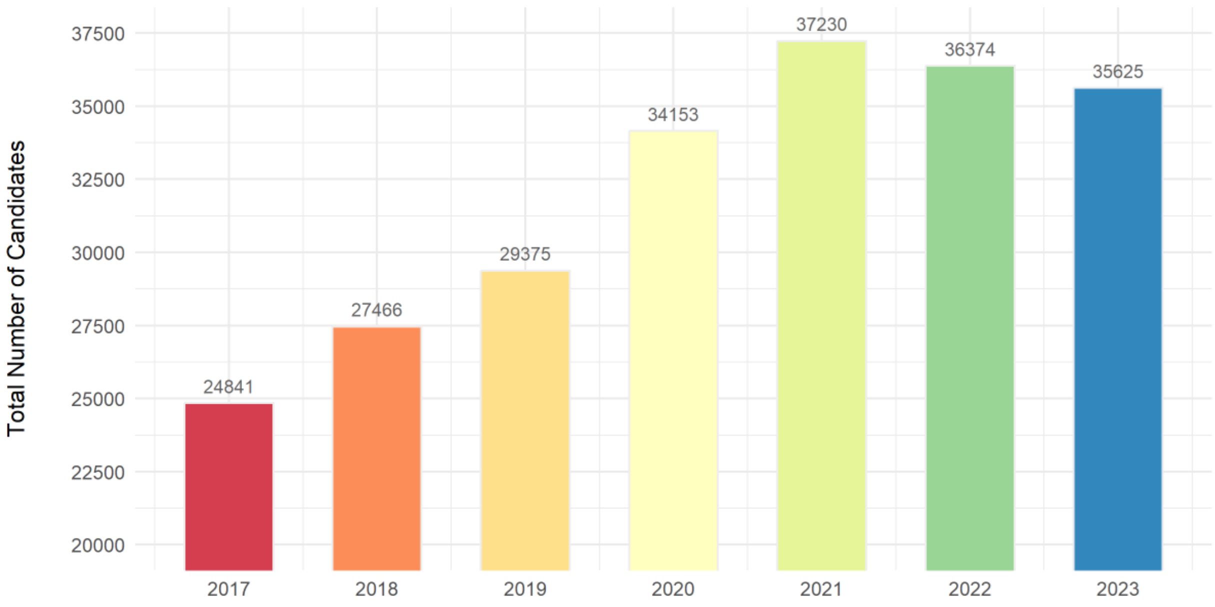


Table 3 presents summary statistics for each of the cognitive subtests plus the total scaled score for the cognitive subtests. VR scores were lowest with a mean score of 591, and the highest average score was achieved on AR with a mean of 652.

Table 3. Cognitive Subtest and Total Scaled Score Summary Statistics

Subtest	Mean	*SD*	Min	Max
VR	590.89	77.94	300	900
DM	623.29	90.35	300	900
QR	649.35	87.13	300	900
AR	652.36	93.51	300	900
Total	2,515.88	286.19	1,320	3,540

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

Over the five-year period, the scaled scores of 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. From 2018 to 2021, AR showed a slow increase trend, and the rest of the subtests were relatively stable. In 2022, a timing adjustment was performed to reduce the speededness of QR, and it was expected that this would result in an increase in the average scaled score. Consequently, QR was scaled down by 20 points to offset this effect. Given the higher scores in AR and QR and lower scores in VR, a decision was made to further decrease the scale score of QR and AR by 10 points each and to increase VR by 20 points in 2023. This rescaling brought the average scaled scores of the subtests closer together in 2023.

Figure 2. Scaled Scores by Year since 2017

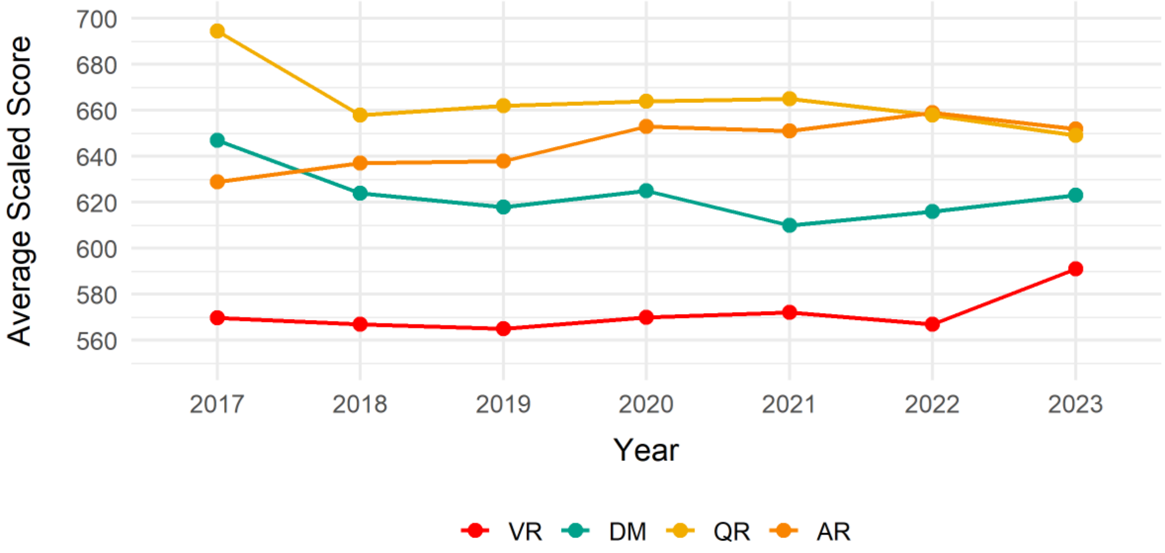


Table 4. Historic Cognitive Subtests Mean Scaled Scores (2017–2023)

Subtest	2017	2018	2019	2020	2021	2022	2023
VR	570	567	565	570	572	567	591
DM	647	624	618	625	610	616	623
QR	695	658	662	664	665	658	649
AR	629	637	638	653	651	659	652

Considering the rescaling efforts detailed earlier, the average performance across subtests have been stable since 2018, with the exception of a small gradual increase in performance in AR. Cohort-to-cohort deviations remain within a few scaled score points after accounting for the rescaling and timing adjustment implemented. These deviations are significantly below one *SEM* for these subtests, as detailed in Section 6. Statistically, these minor deviations are not substantial enough to raise concerns. This stability indicates consistent performance across different cohorts, aligning with expectations given the absence of major test alterations and a stable candidate composition.

All of the subtests have shown a positive significant correlation between each other, indicating that a set of common qualities are measured across all of the subtests, as presented in Table 5.

Table 5. The Scaled Score Zero-Order Correlation of the Subtests

	VR	DM	QR	AR
DM	0.65***			
QR	0.55***	0.70***		
AR	0.37***	0.53***	0.57***	
SJT	0.42***	0.49***	0.45***	0.44***

Note: *** indicates $p < .001$.

For the SJT, the number and percentage of candidates in each band for the 35,625 candidates who took the 2023 UCAT are shown in Table 6 below. Candidates are awarded a band for the SJT exam based on their underlying scaled score.

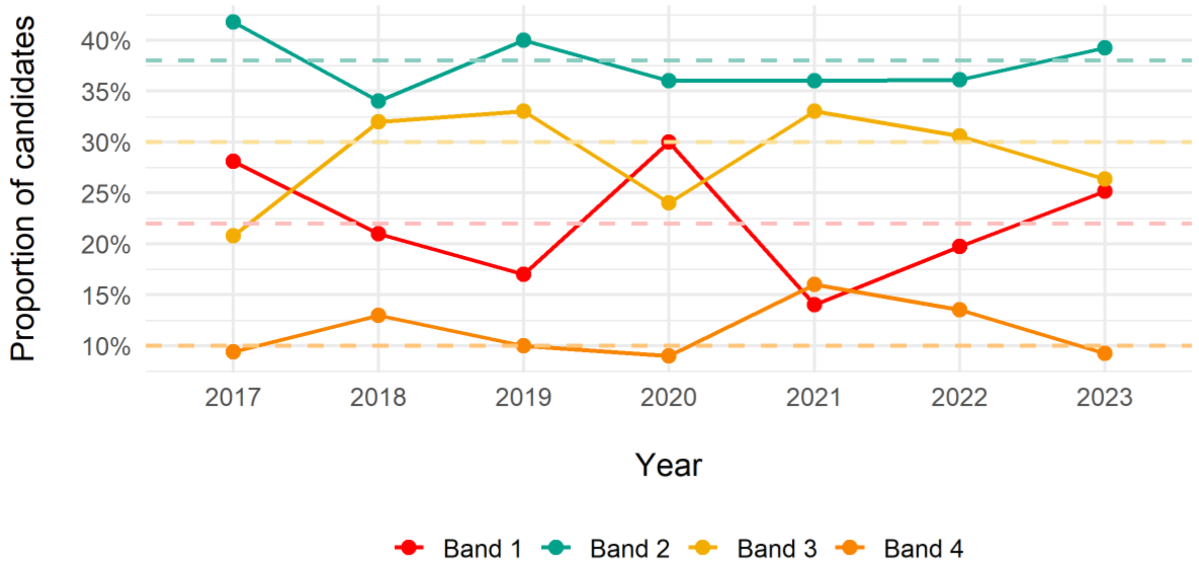
Table 6. SJT Band Distribution in 2023

SJT Band	Number of Candidates	Mean Scaled Score	Percentage of Candidates	Target %
Band 1	8,964	678.97	25%	22%
Band 2	13,981	625.80	39%	38%
Band 3	9,390	553.35	26%	30%
Band 4	3,290	429.32	9%	10%
Total	35,625	601.94	100%	100%

The proportions of candidates in the four different bands deviated from the target. Specifically, the percentages for Bands 1 and 2 exceed the target by three and one percentage points, respectively, while those for Bands 3 and 4 fall short by four and one percentage points, respectively. This shows that the candidates in this cohort performed slightly better than we have anticipated, resulted in a higher proportion of Band 1-2 candidates and a lower proportion of Band 3-4 candidates.

Figure 3 illustrates the distribution of candidates across SJT bands since 2017. From 2018, target proportions for each SJT band were introduced. Although these targets vary annually, they typically fluctuate within a 1% to 2% range. The 2023 target proportions are represented by dotted lines in Figure 3. Generally, the actual proportions align closely with the targets, albeit with minor deviations. This year, the largest significant deviation is 4%, which is consistent with the range of deviation observed in previous years. 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.

Figure 3. SJT Band Proportions 2017–2023



The distribution of scores is important because the band boundaries (defined in Table 2) are set each year by 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 returned to normal, correspondingly fewer candidates were categorised as Band 1. However, the 2022 band thresholds were based on the 2021 population, and therefore the band distributions are much closer to the target. Thresholds for the current year, established on the basis of the 2022 candidate cohort, have yielded a slightly higher proportion of candidates in Bands 1 and 2 than the intended target proportion, suggesting a small increase in the performance of the 2023 cohort relative to that of 2022. These findings will guide the calibration of thresholds for the subsequent year.

4.2 Special Educational Needs

There are five exam versions available for SEN candidates who are allowed extra time and breaks. However, only one candidate undertook the UCATSEN100SA test code during the contingency period, and thus their data is not included in this technical report. Consequently, only four alternative examination versions are reported here. Table 7 below details the time allowances for each subtest and exam version.

Table 7. 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	00:25:00	00:31:15	00:31:15	00:37:30	00:25:00
AR	00:12:00	00:15:00	00:15:00	00:18:00	00:12:00
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, which is consistent with 2022. The most popular SEN exam was UCATSEN, as shown in Table 8 below. These exams are available to candidates who require additional time due to a special accommodation.

Table 8. Exam Version Candidate Volumes

Exam	*N*	%
UCAT	33,631	94%
UCATSEN	1,301	4%
UCATSENSA	422	1%
UCATSEN50	93	0%
UCATSA	178	0%
Total	35,625	100%

Historically, candidates who take a SEN version of the exam usually outperform candidates who take the non-SEN version. Table 9 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 9. SEN and Non-SEN Cognitive Subtests

Subtest	Statistic	UCAT (33,631)	UCATSEN (1,301)	UCATSENSA (422)	UCATSEN50 (93)	UCATSA (178)
VR	Mean	589.35	612.98	631.16	614.73	611.74
	SD	77.58	76.80	86.00	75.68	81.04
	Min	300.00	400.00	390.00	390.00	390.00
	Max	900.00	900.00	900.00	840.00	900.00
DM	Mean	621.81	646.20	656.47	642.58	646.35
	SD	90.19	87.88	98.88	80.69	80.43
	Min	300.00	360.00	320.00	410.00	430.00
	Max	900.00	890.00	890.00	830.00	890.00
QR	Mean	648.25	665.90	674.27	671.94	665.28
	SD	87.27	79.93	87.84	91.05	84.45
	Min	300.00	430.00	360.00	490.00	450.00
	Max	900.00	900.00	900.00	900.00	880.00
AR	Mean	650.94	680.75	669.60	674.95	659.61
	SD	93.76	85.09	86.78	87.24	86.09
	Min	300.00	430.00	300.00	500.00	410.00
	Max	900.00	900.00	900.00	880.00	890.00
Total	Mean	2,510.36	2,605.83	2,631.49	2,604.19	2,582.98
	SD	286.32	259.32	292.17	261.49	265.88
	Min	1,320.00	1,860.00	1,550.00	1,800.00	1,890.00
	Max	3,540.00	3,430.00	3,390.00	3,150.00	3,180.00

Table 9 also includes the mean total cognitive scaled score for each exam version. It is evident that 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 95 scaled score points. This is higher than in 2022, where the difference was 91 scaled score points.

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 by exam version is presented in Table 10 below. The version of the exam on which candidates performed the best is the UCATSA, where 83% of candidates are categorised as either Band 1 or Band 2, but note the prior warning that few candidates sat that version of the exam, meaning comparison may not be reliable.

Table 10. SJT Band by Exam Version

Exam Version	Mean Scaled Score	Band 1	Band 2	Band 3	Band 4
UCAT	600.27	24%	39%	27%	10%
UCATSEN	629.17	36%	42%	19%	3%
UCATSENSA	632.41	36%	46%	15%	3%
UCATSEN50	626.95	33%	43%	20%	3%
UCATSA	633.37	40%	43%	13%	4%

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 11 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 11. 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.

4.3 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 applied to study medicine or dentistry. Candidates who applied for both are categorised according to their first choice.

The majority of candidates applied for medicine, accounting for 59% of candidates, a reduction from 63% in 2022 and 69% in 2021. In contrast, 13% of candidates applied for dentistry, an increase from 11% in 2022 and 9% in 2021. The remaining 29% applied for neither or could not be matched with UCAS data.

Candidates who applied for medicine as a first choice outperformed those who applied for dentistry, as illustrated in Table 12. The highest mean scaled score was achieved on AR and the lowest on VR for both candidate groups. Candidates who did not apply for medicine or dentistry or were not matched by UCAS performed less well than both other groups.

Table 12. Medicine/Dentistry Candidates: Cognitive and Total Scaled Scores

	Mean			*SD*		
Subtest	Medicine	Dentistry	None	Medicine	Dentistry	None
VR	606.69	586.22	561.01	76.62	66.39	76.17
DM	644.31	626.06	579.66	85.89	79.37	88.08
QR	667.96	656.77	608.56	85.15	78.26	80.78
AR	673.26	667.60	603.57	91.47	88.21	81.00
Total	2,592.23	2,536.65	2,352.80	270.51	246.29	264.91

Better performance by medicine candidates is also reflected in the SJT banding. As Table 13 shows, more medicine than dentistry candidates appeared in Band 1, and fewer medicine than dentistry candidates appeared in Band 4.

Table 13. Medicine/Dentistry Candidates: SJT Bands

Group	Mean Scaled Score	Band 1	Band 2	Band 3	Band 4
Dentistry	616.30	29%	44%	22%	5%

Group	Mean Scaled Score	Band 1	Band 2	Band 3	Band 4
Medicine	619.27	31%	43%	22%	4%
None	560.77	12%	30%	37%	21%

In summary, UCAT candidates who applied for medicine performed better across all subtests than those who applied for dentistry and both of these groups performed better than those who applied to neither. This is consistent with test performance in previous years.

4.4 Mode of Delivery

In 2023, the UCAT was offered in both the standard test centre and online proctored mode. Only 31 candidates took the exam in the online proctored mode, amounting to only 0.09% of all candidates. This contrasts with 2020, when more than 11,038 candidates took the exam in the online mode. The proportion of candidates using the online version of the test is decreasing as test centres are back open fully and candidates are encouraged to use a test centre where possible.

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 2023, it is not possible to draw reliable inferences on differences in performance for the 2023 cohort of candidates.

4.5 Examination Results by Demographic Variables

4.5.1 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 that are freely accessible to all. Finally, we analyse the performance of individual items by demographic characteristic and remove any items that might exhibit bias (as discussed in Section 7.3).

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 scaled scores of the cognitive subtests.

4.5.2 Gender

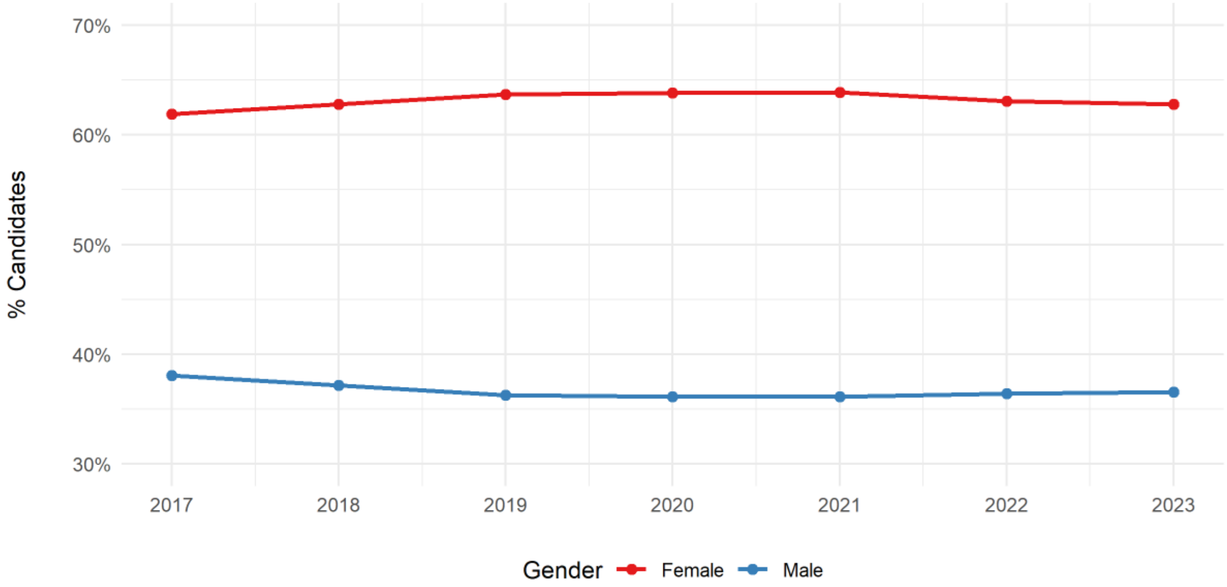
Table 14 presents the breakdown of test-takers by gender. The majority of test-takers were female, and only 240 stated “Other” or that they would prefer not to say.

Table 14. Gender Counts

Gender	*N*	%
Female	22,362	63%
Male	13,023	37%
I prefer not to say	197	1%
Other	43	0%

The distribution of candidates by gender has remained stable since 2017, with a slight increase in female candidates from 2017 to 2019 (Figure 4).

Figure 4. Distribution of Candidates by Gender 2017–2023



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 15, ranging from 10 scaled score points on VR to 33 scaled score points on QR. However, note that these differences are less than the *SEM* on the subtest and therefore may not be significant. Further analysis can be found below.

Table 15. Gender Scaled Scores

Subtest	Mean Scaled Score		*SD* Scaled Score	
	Female	Male	Female	Male
VR	587.01	596.65	77.54	77.73
DM	615.44	636.10	89.41	90.18
QR	636.86	670.42	83.14	89.60
AR	646.70	661.76	91.24	96.37

Subtest	Mean Scaled Score		*SD* Scaled Score	
	Female	Male	Female	Male
Total Cognitive	2,486.01	2,564.93	280.52	288.26
SJT	608.39	590.53	73.16	79.44

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 represents the combined sample sizes of both groups minus two, reflecting independent data points for comparison. A non-zero *t*-statistic indicates there is a difference in the mean scaled 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.

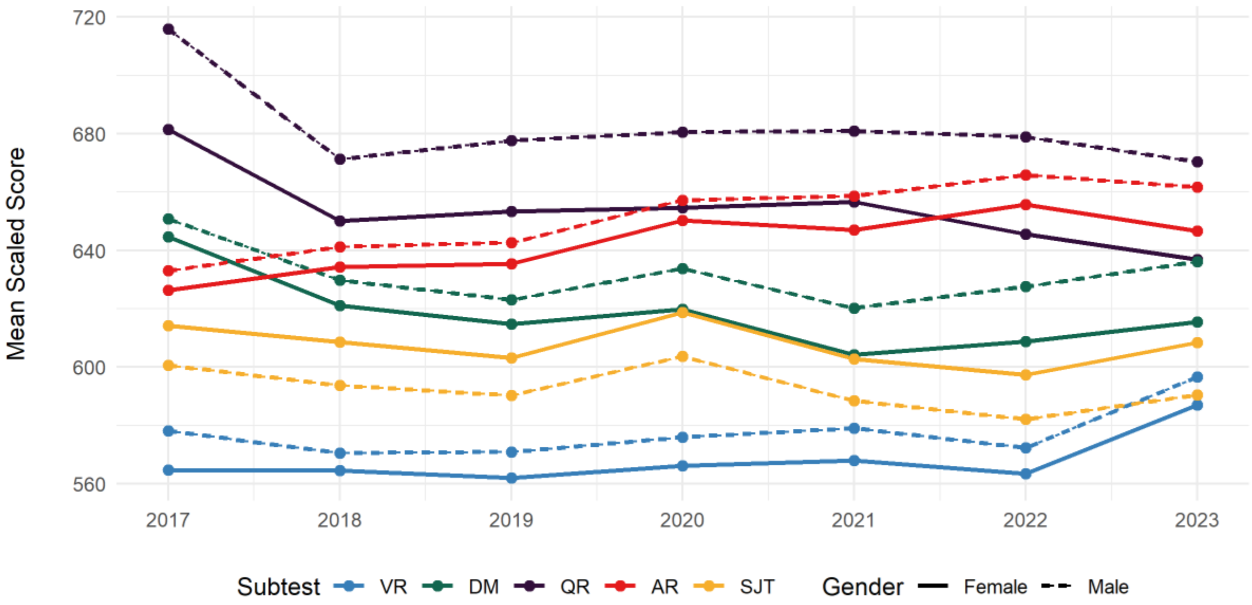
Therefore, Table 16 shows us that there are differences between male and female performance on each subtest and on the total cognitive scores, and that these differences are likely not to be the result of random chance.

Table 16. Gender *t*-Test

Subtest	*t*-Statistic	*df*	*p* Value
VR	11.27	35,383	< 0.01
DM	20.89	35,383	< 0.01
QR	35.58	35,383	< 0.01
AR	14.67	35,383	< 0.01
Total Cognitive	25.26	35,383	< 0.01
SJT	-21.45	35,383	< 0.01

Figure 5 illustrates 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. Since 2021, the range has slightly broadened in the QR subtest.

Figure 5. Scaled Score Distribution of Candidates by Gender 2017–2023



4.5.3 Ethnicity

UCAT candidates who reside in the UK are requested to answer a question relating to their ethnicity. The ethnic categories in the questionnaire were simplified in 2023 by reducing the number of options. These options align closely with the groups used in previous reports except for UK-Chinese, which is no longer a separate category. The categories used are:

- White
- Mixed or multiple ethnic groups
- Asian or Asian British
- Black, African, Caribbean or Black British
- Other ethnic group
- I prefer not to say

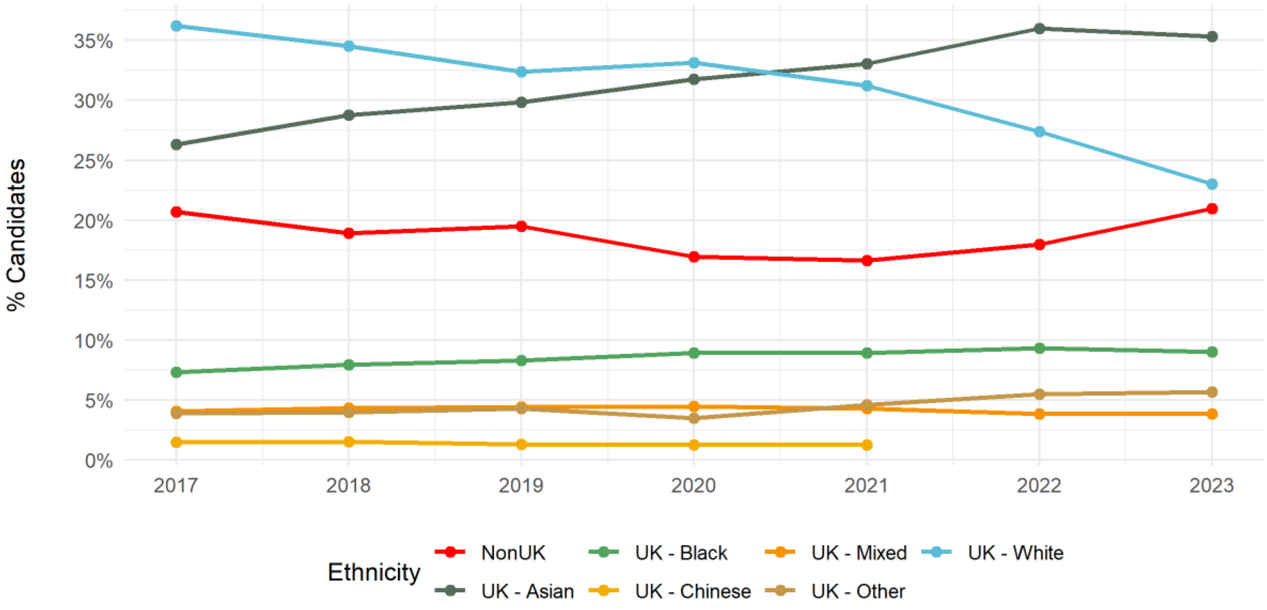
Table 17 shows the breakdown of candidates by ethnicity in the 2023 exam. The biggest candidate group was UK - Asian. Twenty-one percent of candidates were not categorised due to being non-UK candidates.

Table 17. Ethnic Group Counts

Country	Ethnic Group	*N*	% UK Candidates	% Total Candidates
UK	Asian	12,581	46%	36%
UK	White	8,204	30%	24%
UK	Black	3,209	12%	9%
UK	Other ethnic group	2,019	7%	6%
UK	Mixed	1,361	5%	4%
Non-UK	Non-UK	7,466	N/A	21%

The proportion of candidates in each ethnic group has remained fairly stable in recent years. Figure 6 shows that the most common ethnic group changed from White to Asian for the first time in 2021. Since then, the proportion of White candidates has continued to decrease, and candidates who identify as UK - Asian has remained the most common ethnic group. The proportion of non-UK candidates has decreased since 2017 but the proportion of non-UK candidates increased in 2023 to reach a level similar to that in 2017. Note that since 2022, "UK - Chinese" was removed as a possible option on the survey.

Figure 6. Distribution of Candidates by Ethnic Group 2017–2023



UK - White candidates performed better on average on all subtests than other groups. Table 18 shows the average scores in each subtest for each ethnic group. Performance was the lowest for UK - Black candidates on average on all subtests except the SJT, where non-UK candidates received the lowest average scaled scores.

Table 18. Ethnic Group Mean Scaled Score

Subtest	White	Asian	Black	Mixed	Other Ethnic Group	Non-UK
VR	613.41	586.18	571.06	607.94	568.14	585.79
DM	651.49	618.79	588.78	638.28	602.08	618.72
QR	661.27	654.93	612.28	655.79	636.32	646.29
AR	666.61	659.85	621.33	663.67	648.65	637.03
Total Cognitive	2,592.78	2,519.75	2,393.45	2,565.67	2,455.19	2,487.83
SJT	620.62	609.09	596.57	615.31	597.29	569.84

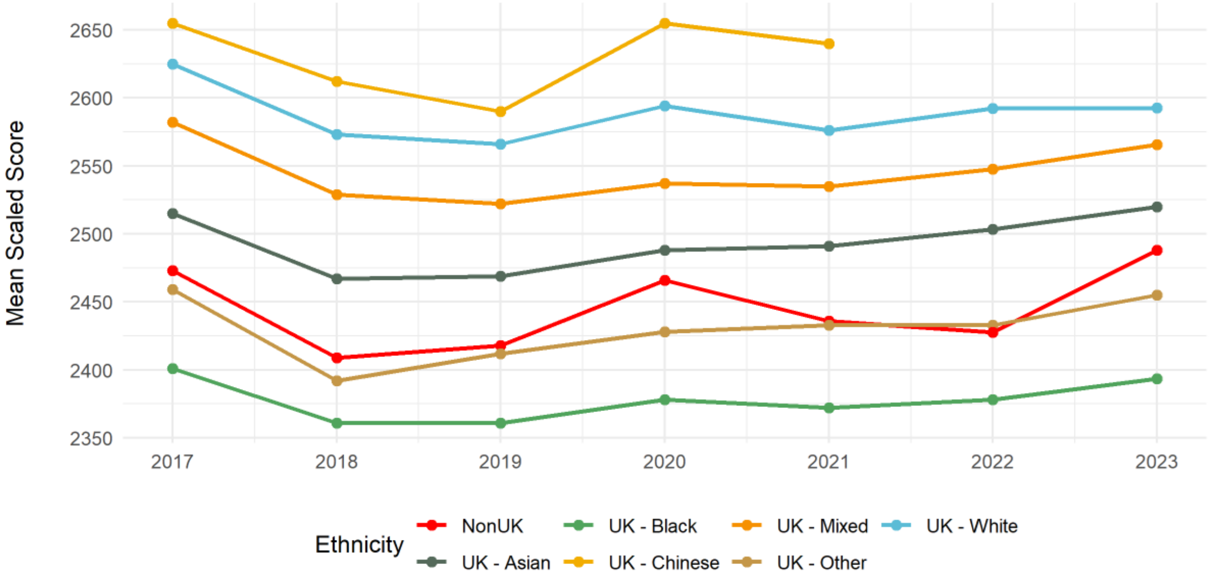
An *F*-test was used to examine whether the differences observed in Table 26 were likely to be due to chance. An *F*-test is similar to the *t*-test discussed in relation to gender (see section 4.5.2). It is used when there are more than two groups. Table 19 has a positive *F*-statistic for each subtest and a *p* value of less than 0.01, which indicates that the differences observed in Table 18 are likely to reflect true differences in performance in the candidate population.

Table 19. Ethnic Group *F*-Test

Subtest	*F*-Statistic	*df*	*p* Value
VR	208.33	6	< 0.01
DM	256.43	6	< 0.01
QR	146.99	6	< 0.01
AR	145.47	6	< 0.01
Total Cognitive	241.65	6	< 0.01
SJT	355.38	6	< 0.01

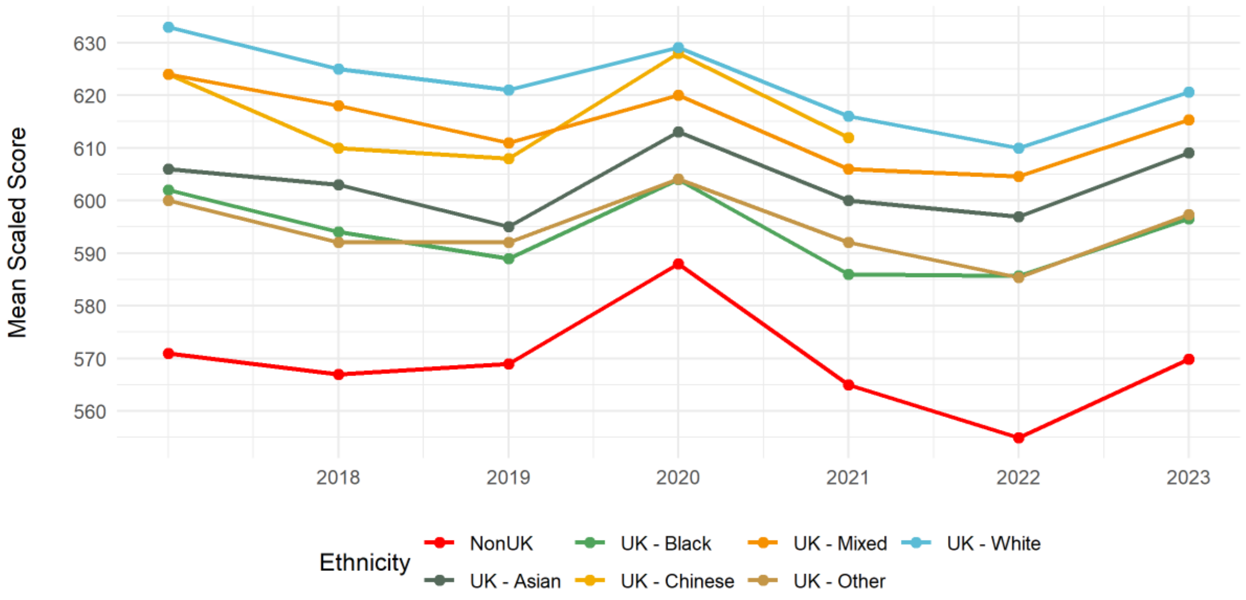
Mean total cognitive scaled scores fell for all ethnic groups between 2017 and 2018 reflecting the rescaling that took place (Figure 7). After 2018, scores have remained fairly stable for most of the ethnic groups, with small increases for Non-UK candidates. The UK - Chinese ethnic category was removed from the survey since 2022.

Figure 7. Ethnic Group Mean Scaled Score for Total Scaled Score 2017–2023



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 2022, with a small increase observed in 2023. The most notable thing about ethnic group trends for the SJT is the margin by which non-UK candidates underperformed relative to the other groups, as can be observed in Figure 8.

Figure 8. Ethnic Group Mean Scaled Score for SJT 2017–2023



The underperformance of non-UK candidates on the SJT might be explained by a link between situational judgement and cultural competence. Specifically, UK-based candidates are more likely to have a better understanding of UK-specific situational norms of behaviour. However, it is important to note that no potential bias against candidates based on residency was identified at item level in the SJT.

4.5.4 Socio-Economic Classification (SEC)

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 if self-employed. Although the primary question about what sort of 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 20, which shows that 23% 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.

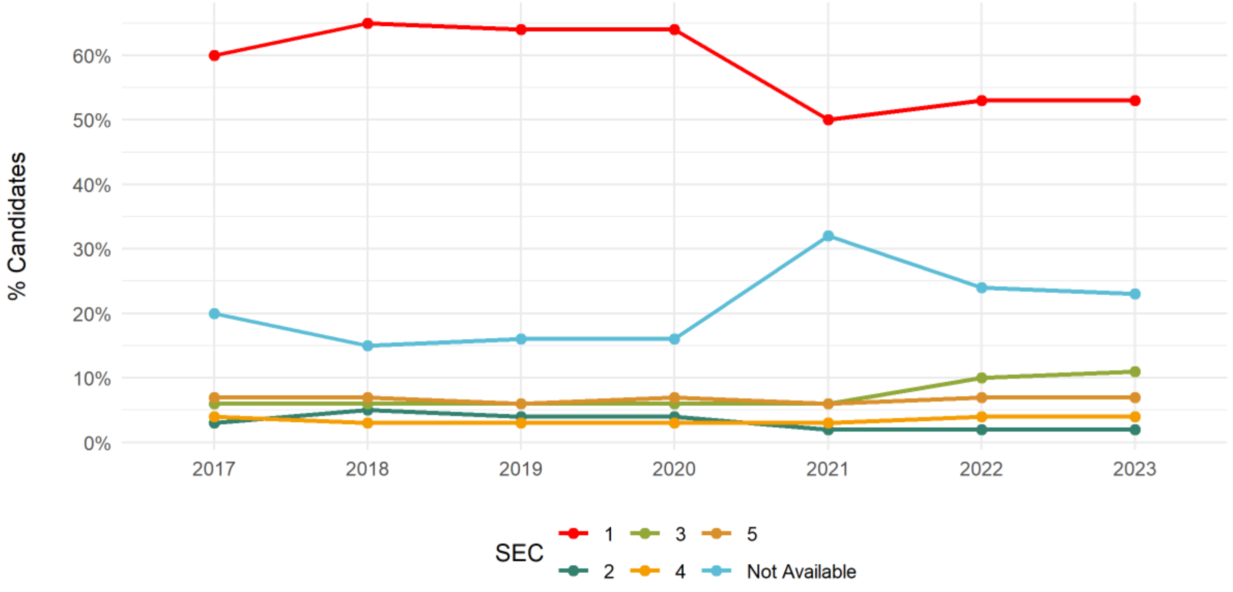
Table 20. SEC Counts

Country	SEC	*N*	% of SEC	% of All
UK	1	14,986	53%	42%
	2	568	2%	2%
	3	3,027	11%	8%
	4	1,051	4%	3%
	5	2,057	7%	6%
	Unknown	6,470	23%	18%
EU		972		3%
Other		6,494		18%

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 from 2021 than in previous years, as illustrated in Figure 9. It suggests that there are fewer candidates in SEC 1 since 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. The trend in 2023 is similar to that observed in 2022.

Figure 9. Candidates by SEC 2017–2023



Consistent with previous years, SEC 1 is the predominant category. Candidates who are SEC 1 also receive higher scores than all other classifications, as shown in Table 21.

Table 21. SEC Scaled Scores

Mean Scaled Score						
Subtest	SEC 1	SEC 2	SEC 3	SEC 4	SEC 5	NA
VR	603.95	591.80	584.31	584.72	574.02	575.88
DM	640.74	618.70	615.15	616.72	599.30	601.07
QR	663.29	637.18	644.40	639.09	628.01	632.42
AR	667.83	644.51	652.44	644.95	637.44	640.80
Total Cognitive	2,575.81	2,492.18	2,496.30	2,485.47	2,438.77	2,450.17
SJT	618.40	612.77	608.23	605.28	602.55	596.24
SD						
VR	75.94	69.59	68.51	71.37	67.82	74.16
DM	86.36	79.97	82.57	81.39	80.09	88.20
QR	83.73	78.66	79.05	77.35	76.83	82.96
AR	92.73	82.89	89.10	88.07	88.96	91.89
Total Cognitive	272.72	248.06	258.07	252.89	251.84	277.30
SJT	64.12	66.40	67.62	68.42	71.28	76.63

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 22 shows that the score differences observed in each subtest are likely to be due to true differences.

Table 22. SEC *F*-Test

Subtest	*F*-Statistic	*df*	*p* Value
VR	155.56	5	< 0.01
DM	208.71	5	< 0.01
QR	144.42	5	< 0.01
AR	156.72	5	< 0.01
Total Cognitive	246.00	5	< 0.01
SJT	354.67	5	< 0.01

4.5.5 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 23). A steady proportional increase in candidates aged 16–19 taking the test can be observed; 76% of the testing population was aged 16–19 in 2020, 78% in 2021, 81% in 2022 and 82% in 2023.

Table 23. Age Counts

Age	*N*	Percent
<= 15	77	0%
16–19	29,181	82%
20–24	4,754	13%
25–34	1,325	4%
>= 35	281	1%

Candidates who were aged 16–19 tended to perform better in all cognitive subtests, as illustrated in Figure 10 below. In the SJT, candidates who were 20–24 tended to perform the best. Candidates who were under 16 and over 34 typically had the lowest performance on the exam; however, the small group sizes for those categories means it is difficult to draw meaningful conclusions from that information. Overall, candidates who were aged 16–19 performed better than other candidates when evaluated by their total cognitive scaled scores, followed by the candidates who were aged 20–24, as illustrated in Figure 11.

Figure 10. Mean Scaled Scores by Age

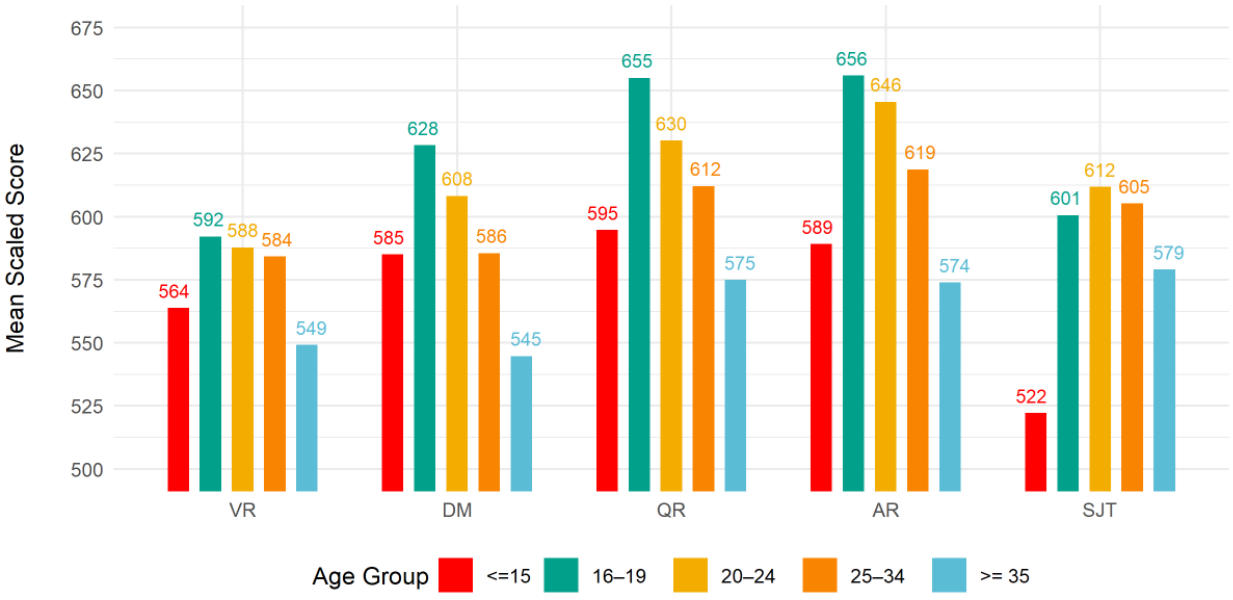
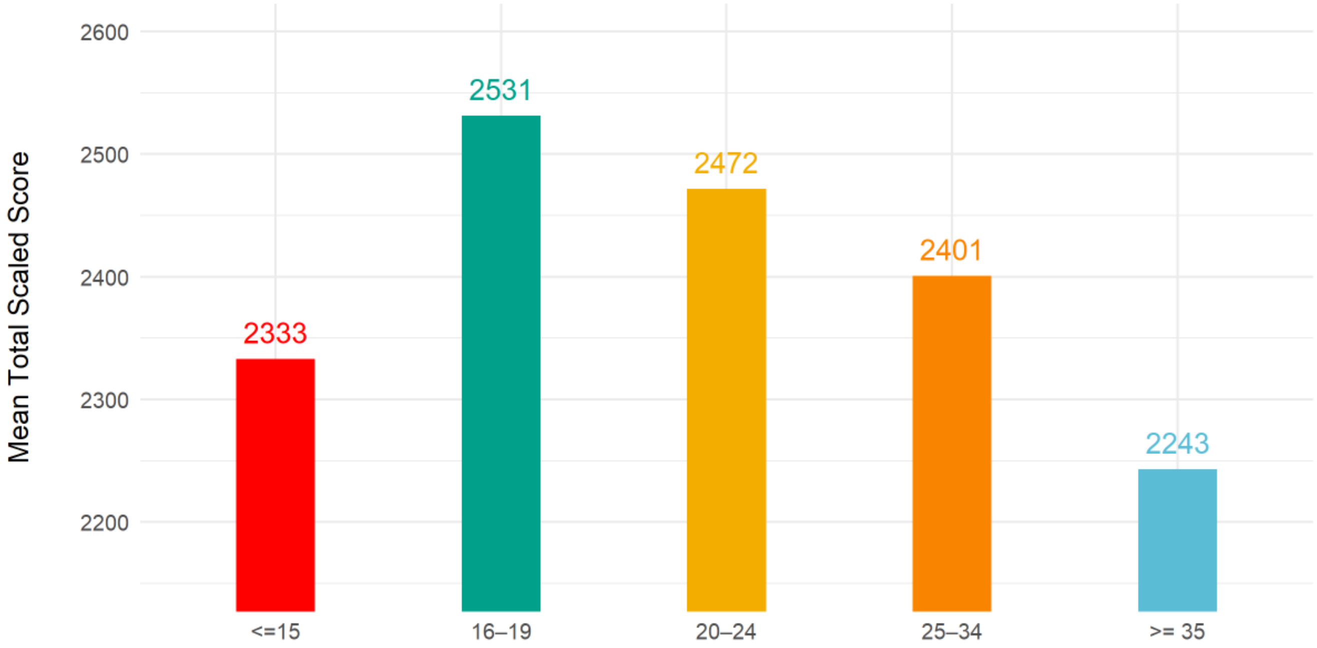


Figure 11. Mean Total Scaled Scores of Cognitive Subtests by Age



Hypothesis testing demonstrated that the differences observed among the groups is unlikely to have occurred due to chance, as shown in Table 32.

Table 24. Age *F*-Test

Subtest	*F*-Statistic	*df*	*p* Value
VR	28.60	4	< 0.01
DM	173.27	4	< 0.01
QR	213.01	4	< 0.01
AR	119.27	4	< 0.01
Total	178.67	4	< 0.01
SJT	51.08	4	< 0.01

To understand how age relates to subtest performance, Table 25 shows the correlation between candidate age and their performance on each subtest. As the significance column shows, all the subtests had statistically significant correlations except for the SJT. 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 negative correlation is for QR. No significant correlation between age and SJT subtest was observed for the year 2023.

Table 25. Correlation of Scaled Score with Age (ungrouped)

Subtest	Correlation	Significance
VR	-0.06	*p* < 0.01
DM	-0.14	*p* < 0.01
QR	-0.15	*p* < 0.01

Subtest	Correlation	Significance
AR	-0.11	$p < 0.01$
Total Cognitive	-0.14	$p < 0.01$
SJT	0.01	$p = 0.20$

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

4.5.6 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 candidates in 2023 had a school leaver qualification (84%), 15% had a degree or above (down from 16% in 2022), and a small minority had no formal qualifications.

Candidates with a degree or above performed better on average on the SJT. For the cognitive subtests and the total cognitive score, below-honours degree candidates performed better on average, as shown in Table 26.

Table 27 shows that the differences observed in Table 26 are statistically significant.

Table 26. Education Scaled Scores

Subtest	School Leaver Qualification	Degree Level or Above
Mean Scaled Score		
N	29,850	5,178
VR	592.21	587.53
DM	627.61	602.99
QR	654.19	625.60
AR	655.48	638.39
Total Cognitive	2529.48	2454.51
SJT	601.32	611.68
SD		
VR	77.22	80.56
DM	89.90	88.97
QR	87.55	79.59
AR	93.06	93.44
Total Cognitive	284.76	280.21
SJT	74.97	75.13

Table 27. Education *t*-Test

Subtest	*t*-Statistic	*df*	*p* Value
VR	-4.00	35,026	< 0.01
DM	-18.22	35,026	< 0.01
QR	-21.97	35,026	< 0.01
AR	-12.19	35,026	< 0.01
Total Cognitive	-17.53	35,026	< 0.01
SJT	9.17	35,026	< 0.01

4.5.7 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 reside in the UK, as can be seen in Table 28 below.

Table 28. Candidate Count by Residence

Country of Permanent Residence	*N*	Percent
UK	28,159	79%
Rest of World	6,494	18%
EU	972	3%

As 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 been relatively stable, as shown in Figure 12 below.

Figure 12. Country of Residence 2017–2023

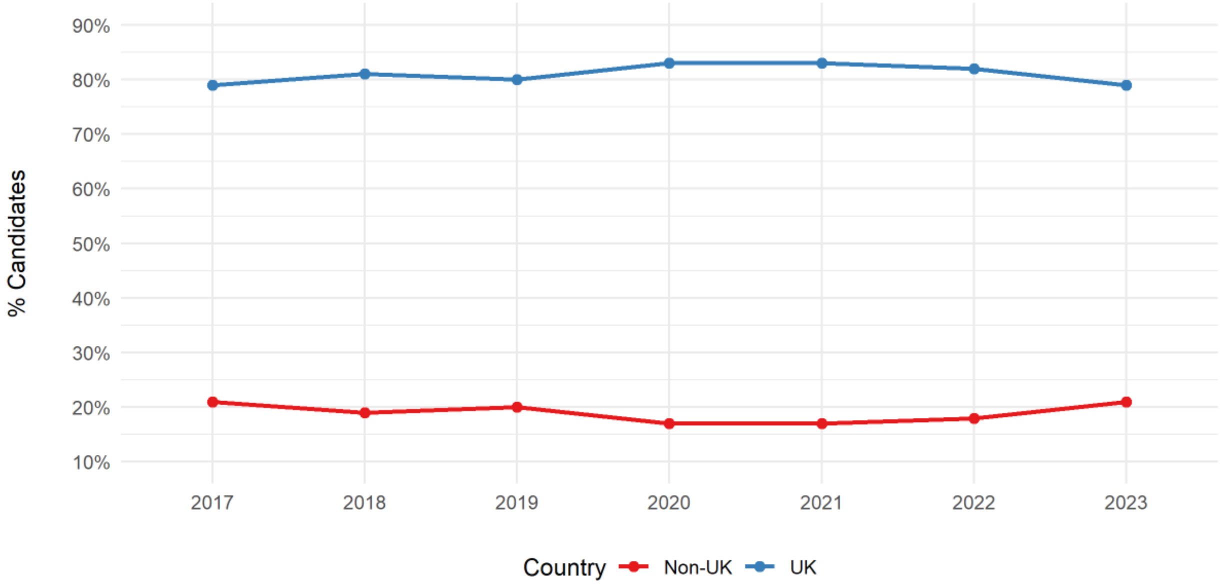


Table 29 shows that UK candidates outperform EU and Rest of World candidates across all subtests, except for QR, in which the Rest of World candidates showed a stronger performance than both the UK and EU candidates.

Table 29. Candidate Scaled Scores by Residence

Subtest	UK	Rest of World	EU
Mean Scaled Score			
VR	592.24	586.20	583.05
DM	624.51	620.55	606.47
QR	650.16	650.34	619.25
AR	656.42	637.86	631.51
Total Cognitive	2,523.32	2,494.95	2,440.28
SJT	610.45	567.58	584.92
SD			
VR	75.02	89.58	75.68
DM	87.51	102.38	83.46
QR	83.51	102.10	75.08
AR	92.39	97.12	88.86
Total Cognitive	275.73	329.08	256.53
SJT	68.91	92.69	79.05

An *F*-test of the differences observed between UK and non-UK candidates is presented in Table 30 below. It shows that the differences are statistically significant.

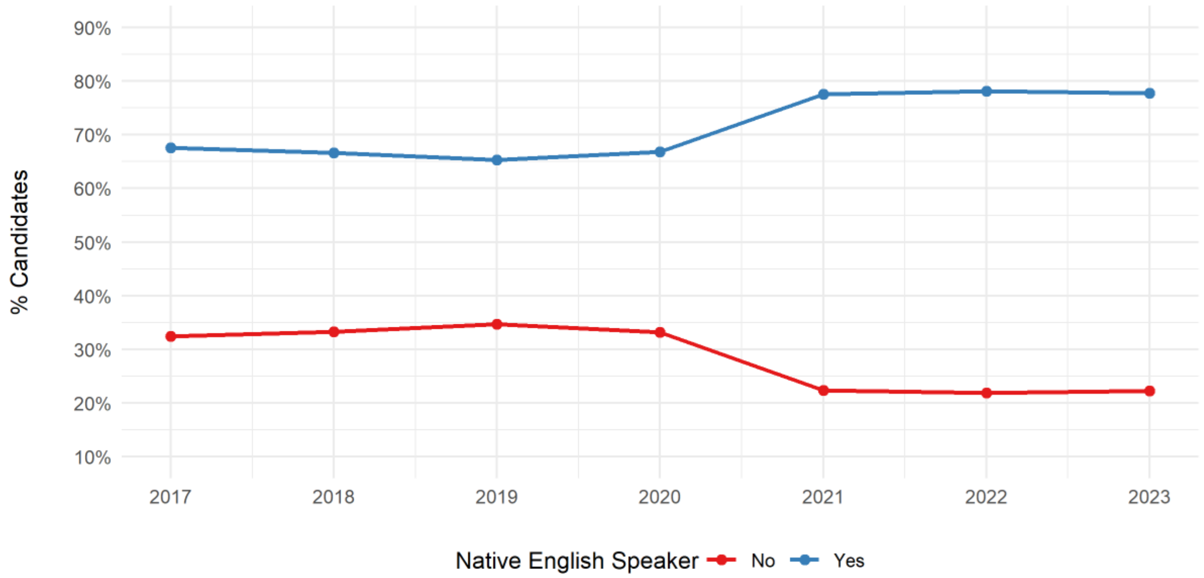
Table 30. Residence *F*-Test

Subtest	*F*-Statistic	*df*	*p* Value
VR	20.94	2	< 0.01
DM	22.40	2	< 0.01
QR	59.83	2	< 0.01
AR	129.72	2	< 0.01
Total Cognitive	61.02	2	< 0.01
SJT	909.85	2	< 0.01

4.5.8 First Language

In 2023, most candidates who sat the UCAT stated that English was 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 (Figure 13). However, between 2022 and 2023 the proportion of candidates with English as a first language stayed at 78%. The change in 2021 is due to a small change in the wording of this question.

Figure 13. Count of Language 2017–2023



Across all subtests, candidates who stated that English was their first language outperformed those who stated that English was not their first language regardless of their country of residence, as shown in Table 31 below.

Table 31. Scaled Scores by Language and Country of Residence

Subtest	Country of Residence	First Language	*N*	% of *N*	Mean	*SD*
VR	UK	English	23,141	65%	599.37	73.93
		Other	5,018	14%	559.36	71.15
	non-UK	English	4,548	13%	609.67	86.07
		Other	2,918	8%	548.56	77.13
DM	UK	English	23,141	65%	632.85	85.52
		Other	5,018	14%	586.04	86.30
	non-UK	English	4,548	13%	641.01	96.50
		Other	2,918	8%	583.97	95.96
QR	UK	English	23,141	65%	655.89	82.30
		Other	5,018	14%	623.71	83.97
	non-UK	English	4,548	13%	663.25	97.70
		Other	2,918	8%	619.85	96.61
AR	UK	English	23,141	65%	660.61	91.99
		Other	5,018	14%	637.08	91.75
	non-UK	English	4,548	13%	644.99	93.61
		Other	2,918	8%	624.63	98.61
Total Cognitive	UK	English	23,141	65%	2548.72	269.00
		Other	5,018	14%	2406.20	276.15
	non-UK	English	4,548	13%	2558.92	308.07
		Other	2,918	8%	2377.02	309.38
SJT	UK	English	23,141	65%	614.87	65.16

Subtest	Country of Residence	First Language	*N*	% of *N*	Mean	*SD*
		Other	5,018	14%	590.09	81.02
	non-UK	English	4,548	13%	588.00	76.40
		Other	2,918	8%	541.53	104.28

In line with the other demographic categories, a test was carried out to understand whether the differences observed in Table 31 can be considered true reflections of the differences between the two groups. Table 32 shows that that such differences are unlikely to have occurred by chance.

Table 32. Language *t*-Test

Subtest	*t*-Statistic	*df*	*p* Value
VR	47.45	35,623	< 0.01
DM	43.63	35,623	< 0.01
QR	31.82	35,623	< 0.01
AR	21.59	35,623	< 0.01
Total Cognitive	43.64	35,623	< 0.01
SJT	40.41	35,623	< 0.01

4.5.9 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, testing 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.

A shortened version of that analysis was also conducted this year to continue monitoring the differences in the performance between UCAT candidates and UCATSEN candidates, as presented in Table 33. After controlling for the effect of the demographic variables (see the note in Table 33), the difference in exam version still explains a significant amount of variance in the candidates' performance, as candidates who took the UCATSEN performed better than those who took the UCAT. The largest difference was observed in the AR subtest, and the smallest difference was observed in the QR subtest. In 2022, the largest difference observed was in QR and the smallest was in SJT, which correspond to the most speeded and least speeded subtests of the exam respectively. The pattern in 2022 had led to the hypothesis that the SEN exam advantage is positively associated with the speededness of the exam. This year results are contradicting to this hypothesis, as both QR and AR are relatively speeded subtests. The performance differences between UCAT and UCATSEN will be continuously monitored in future years to ensure test fairness to all candidates.

Table 33. Subtest Performance Differences: UCAT and UCATSEN (controlling for demographic variables)

Subtest	*F*	*p*	η^2
VR	99.43	<.0001	0.0026
DM	109.43	<.0001	0.0029
QR	75.79	<.0001	0.0020
AR	128.47	<.0001	0.0035
SJT	98.25	<.0001	0.0026

Note. The comparison was only made between UCAT and UCATSEN exam codes, which accounted for 99% of the candidates. The rest of the accommodated exam codes were not included because of the small number of candidates. The demographic variables that were controlled included gender, SEC, age group, highest academic qualification, country of residence and first language. Candidates' ethnicity was not included in the analysis as more than 20% of candidates did not provide this information.

Despite the consistent differences observed in the SEN exam across the years, the effect size, eta-squared η^2, of these differences across all subtests is less than 0.005 after controlling for the effect of the demographic variables, indicating the effect sizes of the differences are very small. The small effect size suggests that the performance gap is not worryingly large considering the normal variation in participants' performance after accounting for the differences in candidates' demographic composition.

5. Exam Timing Analysis

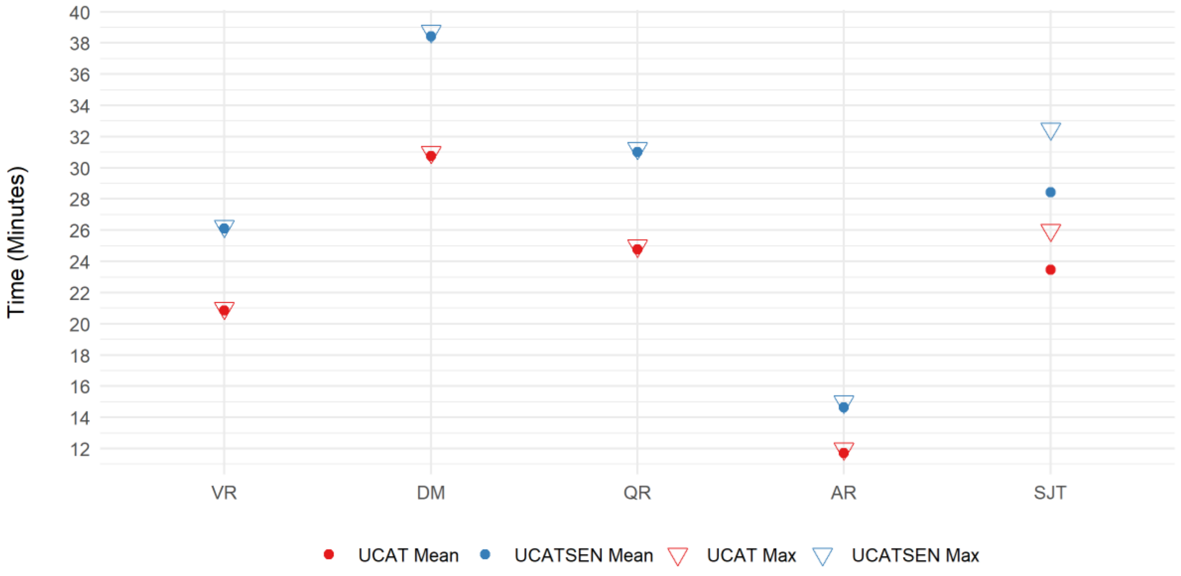
The section time for each candidate is calculated by summing the item and review time for each item and candidate. Table 34 shows the exam timing for each version of the UCAT.

Table 34. Mean Subtest Section Timing: Non-SEN and SEN

Statistic	Subtest	UCAT (33631)	UCATSEN (1301)	UCATSENSA (422)	UCATSEN50 (93)	UCATSA (178)
Mean	VR	00:20:52	00:26:06	00:26:04	00:31:19	00:20:54
	DM	00:30:45	00:38:26	00:38:25	00:45:57	00:30:46
	QR	00:24:46	00:30:60	00:30:56	00:37:15	00:24:49
	AR	00:11:42	00:14:38	00:14:33	00:17:24	00:11:40
	SJT	00:23:28	00:28:25	00:27:21	00:30:32	00:23:06
SD	VR	00:00:27	00:00:23	00:00:35	00:00:23	00:00:10
	DM	00:00:58	00:00:60	00:00:47	00:01:20	00:00:35
	QR	00:01:08	00:01:10	00:01:22	00:00:37	00:00:28
	AR	00:00:47	00:00:54	00:01:04	00:01:17	00:00:50
	SJT	00:03:31	00:05:13	00:05:51	00:07:53	00:03:50
Min	VR	00:01:38	00:19:02	00:18:42	00:28:51	00:19:50
	DM	00:03:03	00:23:07	00:31:58	00:38:52	00:26:34
	QR	00:01:08	00:08:42	00:08:45	00:34:07	00:21:20
	AR	00:00:45	00:05:56	00:05:09	00:10:39	00:06:58
	SJT	00:00:51	00:09:18	00:11:53	00:13:23	00:08:14
Max	VR	00:21:00	00:26:15	00:26:15	00:31:31	00:21:00
	DM	00:31:00	00:38:45	00:38:45	00:46:31	00:31:00
	QR	00:25:00	00:31:15	00:31:15	00:37:31	00:25:00
	AR	00:12:00	00:15:00	00:15:00	00:18:01	00:12:00
	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 presented in Table 34. The cognitive subtests on the UCAT version of the exam are quite speeded. The mean 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 shown in Figure 14 below. The difference between the average time and the maximum time allowed is barely observable for VR and QR for both UCAT and UCATSEN. The difference is slightly broader for DM and 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 35. It shows that the most speeded non-SEN subtests are VR and QR, where 87% and 87% of candidates respectively reached all the items and between 6% to 7% of candidates did not reach five or more items. The SJT is the least speeded in all exam versions.

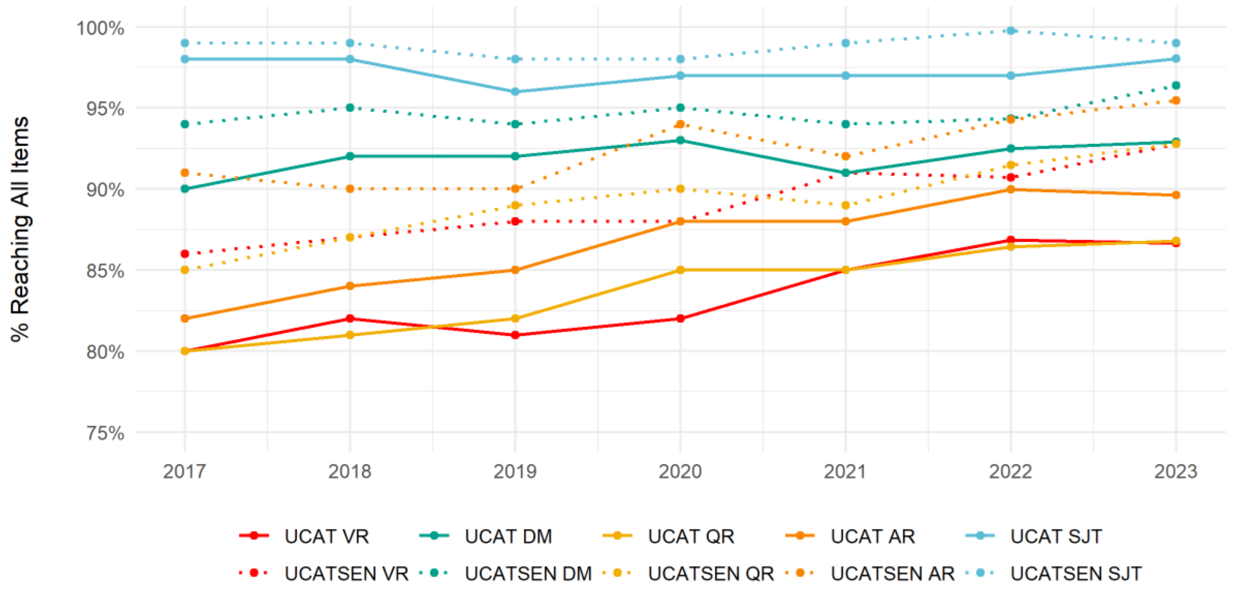
Table 35. Subtest Section Timing: Non-SEN and SEN UCAT Incomplete Tests

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	29,148	87%	2,283	7%	6.78 (4483)
	DM	31,245	93%	719	2%	3.57 (2386)
	QR	29,183	87%	2,101	6%	6 (4448)
	AR	30,144	90%	1,638	5%	6.61 (3487)
	SJT	32,973	98%	118	0%	3.66 (658)
UCATSEN	VR	1,207	93%	39	3%	5.36 (94)
	DM	1,254	96%	7	1%	2.53 (47)
	QR	1,207	93%	40	3%	5.32 (94)
	AR	1,242	95%	17	1%	4.25 (59)
	SJT	1,288	99%	1	0%	2.23 (13)
UCATSENSA	VR	388	92%	16	4%	6.53 (34)
	DM	402	95%	4	1%	3.35 (20)
	QR	386	91%	19	5%	6.14 (36)
	AR	384	91%	19	5%	7.05 (38)
	SJT	416	99%	1	0%	4 (6)
UCATSEN50	VR	86	92%	3	3%	4.29 (7)
	DM	91	98%	0	0%	2.5 (2)
	QR	89	96%	1	1%	2.75 (4)

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
	AR	90	97%	1	1%	4.33 (3)
	SJT	93	100%	0	0%	N/A
UCATSA	VR	159	89%	8	4%	5 (19)
	DM	169	95%	2	1%	2.67 (9)
	QR	157	88%	11	6%	6.1 (21)
	AR	166	93%	6	3%	7.17 (12)
	SJT	177	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 from 2017 to 2021.

Figure 15. Candidates Reaching All Items 2017–2023



In 2022, a change was made to the timing of the AR and QR subtests with the aim of reducing the speededness of QR. One minute was taken from the AR subtest (with the removal of 5 pretest items) and this was added to the QR subtest (where no additional items were included). The item time has been considered in the form build for QR and AR for a number of years, but this was also extended to VR and DM in 2022. A notable increase in the percentage of candidates reaching all items has been observed since 2022. There are no major changes regarding test speededness in 2023 and the percentages of candidates reaching all items are similar to those in 2022.

The factor of guessing has taken into account when evaluating speededness since 2022. Figure 16 to Figure 20 illustrate the distribution of item time usage for the five subtests. With a considerable sample size, these distributions are theoretically expected to conform to a bell-shaped curve as per the central limit theorem. However, bimodal distributions observed in the VR, DM, and QR subtests suggest the presence of two distinct behavioural patterns. The left-hand side peak (local maximum) of these distributions, centred around 2 seconds with a narrow spread, contrasts with the broader peak (local maximum) on the right-hand side. This left-hand peak likely signifies a pattern of rushed guessing behaviour, as it is implausible for any distinct item type to be completed in such a short time span. The right-hand peak, conversely, appears to represent the actual time spent on non-guessed items. The valley (local minimum) between these peaks marks the overlap of the two distributions. By excluding responses shorter than the valley duration, it is probable that a majority of guessed responses, along with some swiftly answered non-guessed responses, can be filtered out. This method offers a feasible approach for estimating speededness for the VR, DM, and QR subtests, by discounting guessed responses. In contrast, the AR and SJT subtests show skewed unimodal distributions, presumably due to low item response times intermingling with guessed responses. This pattern complicates the assessment of speededness for these subtests, as excluding responses below a certain threshold may not effectively separate guessed from non-guessed responses.

Figure 16. VR Item Time Distribution

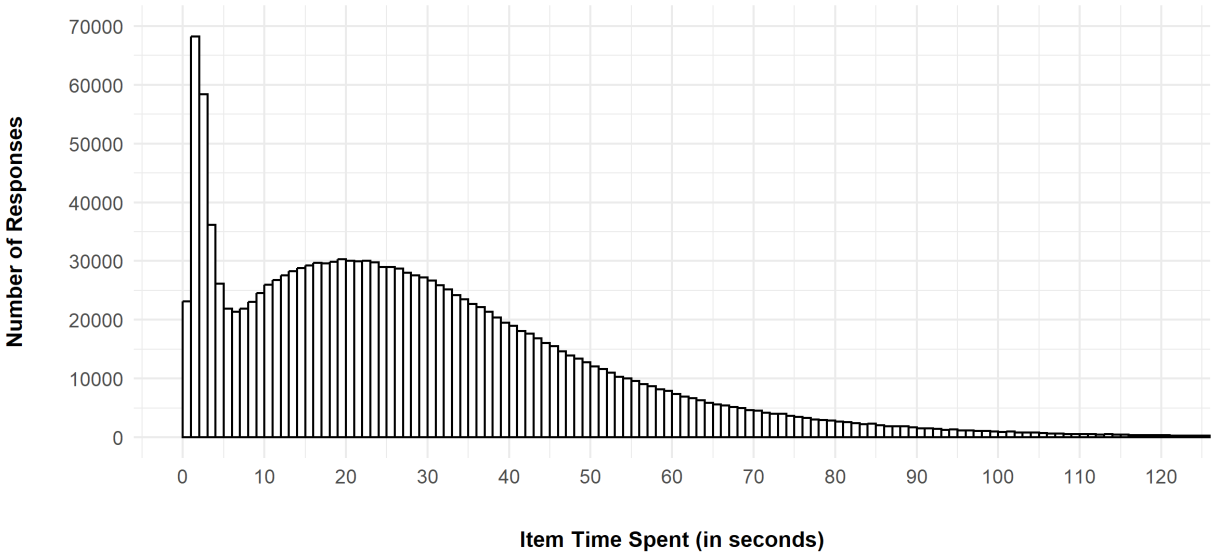


Figure 17. DM Item Time Distribution

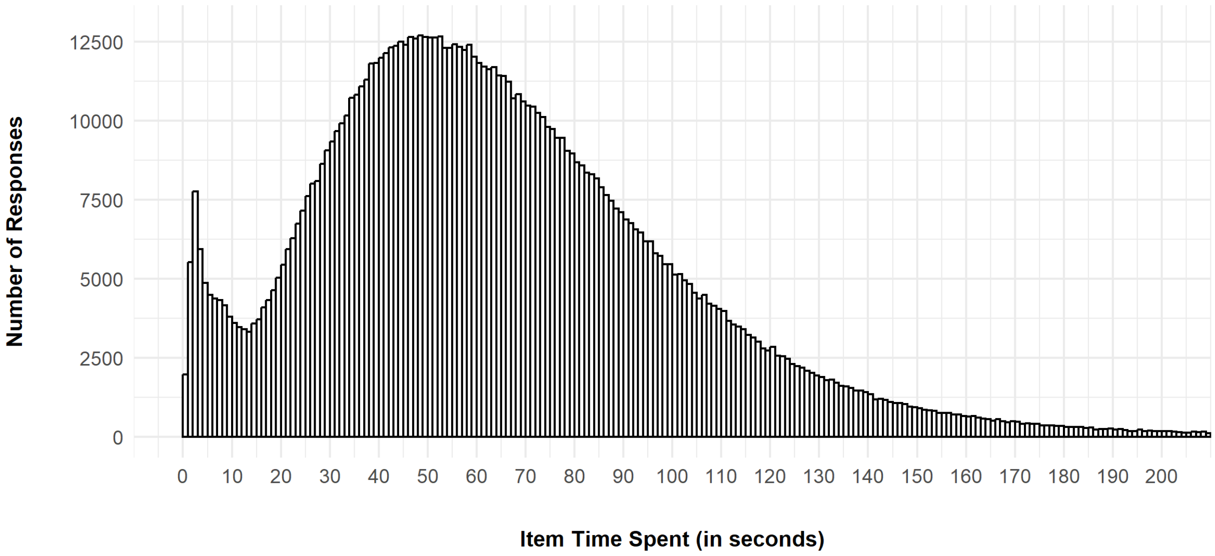


Figure 18. QR Item Time Distribution

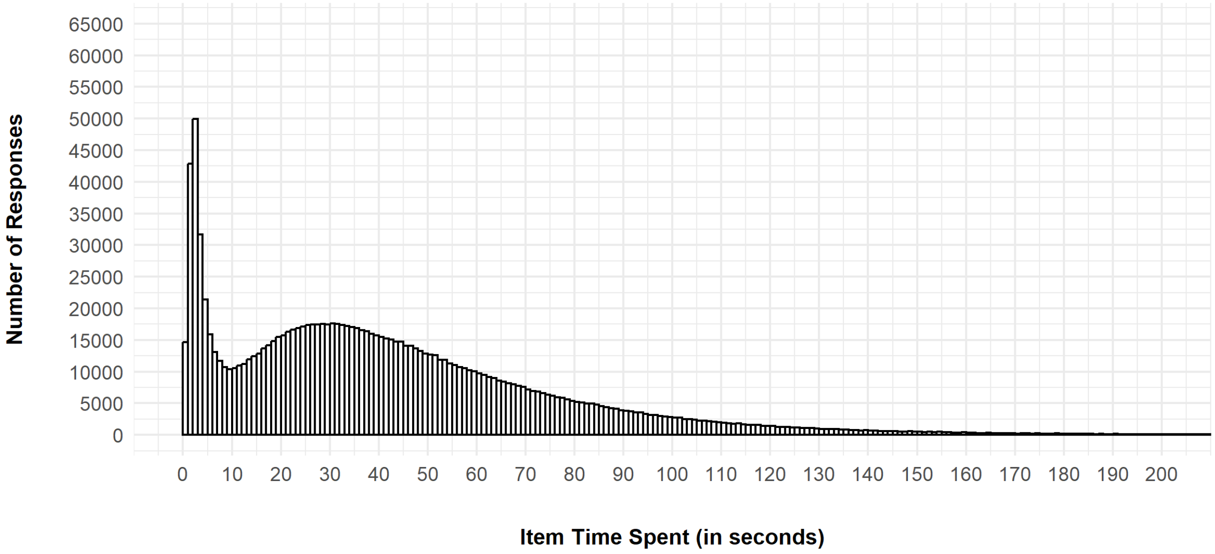


Figure 19. AR Item Time Distribution

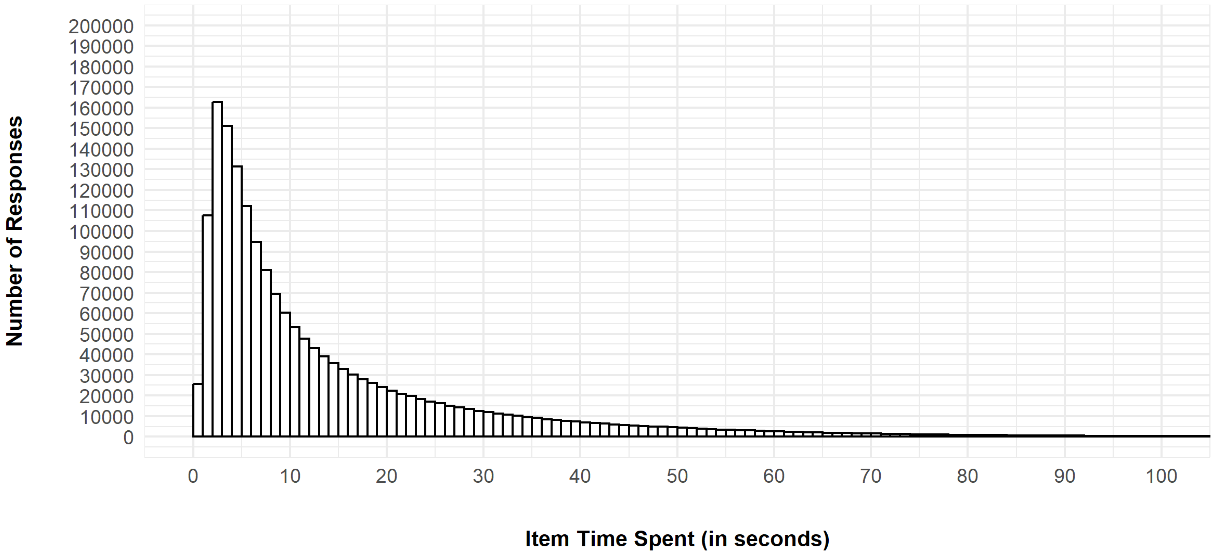
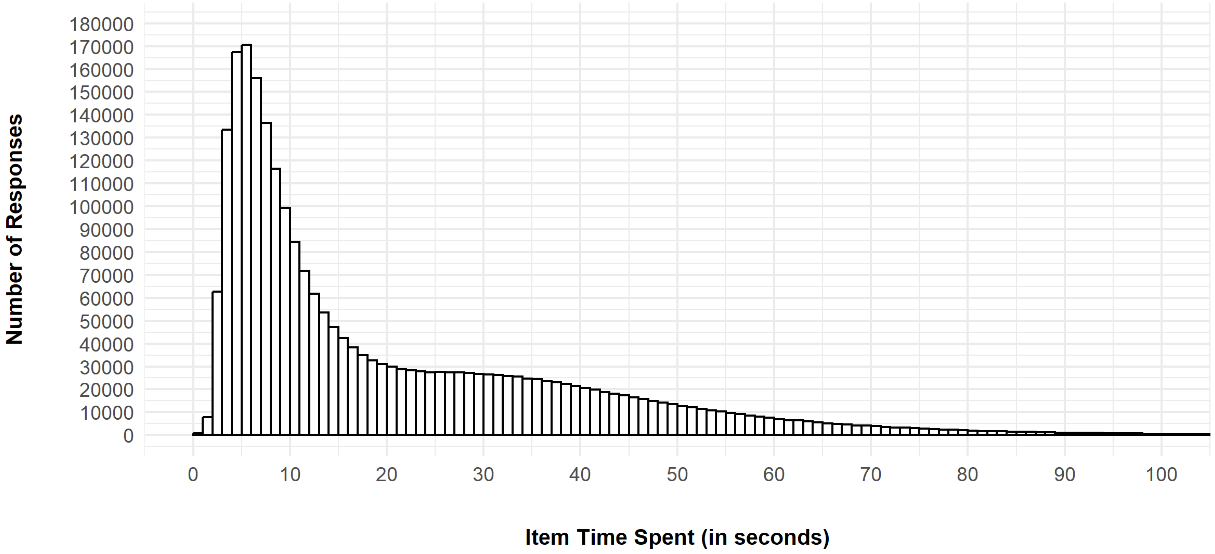


Figure 20. SJT Item Time Distribution



The further examination of speededness for the VR, DM, and QR subtests involved excluding responses based on various guessing thresholds. The threshold for exclusion is a relatively subjective decision that would yield different results. A 1-second threshold, used in previous years, predominantly excluded only the most hasty responses; a 5-second threshold effectively removed the peak and those below the peak of the guessing distribution, eliminating most guessed responses and a minor portion of overlapping non-guessed responses; a 10-second threshold, surpassing the valley for both VR and QR and approximating that of DM, likely filtered out nearly all guessed responses but also removed a significant number of non-guessed responses.

Although a similar analysis was conducted for AR and the SJT, it serves primarily for comparative purposes only. Due to the overlapping distributions of guessed and non-guessed responses in these subtests, as previously discussed, applying a fixed threshold is less effective and could inadvertently exclude a substantial number of non-guessed responses. Consequently, the results for AR and the SJT, detailed in Table 36, should be interpreted with caution.

Assuming a balanced approach of a 5-second exclusion threshold, the proportion of candidates completing all items in VR, DM, and QR without guessing dropped dramatically to 14%, 68%, and 20%, respectively. This indicates that only a small fraction of candidates were able to finish these subtests within the allotted time without resorting to guessing. However, candidates on average reached 84%, 97%, and 85% of the items in VR, DM, and QR, respectively. This suggests that while many candidates could not complete every item without guessing, they were typically able to answer most items in the subtests. Regardless of the guessing exclusion, VR and QR remained the most speeded subtests, with VR being marginally more speeded than QR.

Table 36. Proportion of Test Reached After Guessing Responses Excluded

Subtest	Guessing Threshold	% Candidates Reached All Items	% of the subtest reached			
			Mean	Q1	Median	Q3
VR	All responses included	87%	98%	100%	100%	100%
	Excluding responses ≤ 1s	69%	97%	98%	100%	100%
	Excluding responses ≤ 5s	14%	84%	77%	86%	95%
	Excluding responses ≤ 10s	1%	77%	68%	80%	86%
DM	All responses included	93%	99%	100%	100%	100%
	Excluding responses ≤ 1s	90%	99%	100%	100%	100%
	Excluding responses ≤ 5s	68%	97%	97%	100%	100%
	Excluding responses ≤ 10s	51%	95%	93%	100%	100%
QR	All responses included	87%	98%	100%	100%	100%
	Excluding responses ≤ 1s	75%	97%	97%	100%	100%
	Excluding responses ≤ 5s	20%	85%	78%	89%	97%
	Excluding responses ≤ 10s	9%	81%	69%	83%	92%
AR and SJT results for reference only						
AR	All responses included	90%	99%	100%	100%	100%
	Excluding responses ≤ 1s	72%	97%	98%	100%	100%
	Excluding responses ≤ 5s	0%	66%	58%	66%	74%
	Excluding responses ≤ 10s	0%	43%	36%	42%	48%
SJT	All responses included	98%	100%	100%	100%	100%
	Excluding responses ≤ 1s	98%	100%	100%	100%	100%
	Excluding responses ≤ 5s	2%	85%	80%	87%	93%
	Excluding responses ≤ 10s	0%	57%	49%	58%	65%

6. Test Form Analysis

The 2023 UCAT consisted of five test forms, Table 37 shows the number of candidates who received each form.

Table 37. Candidates by Form

Form	Candidates
Form 1	46
Form 2	9,415
Form 3	9,404
Form 4	8,428
Form 5	8,332

Table 38 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. All subtests have satisfactorily high reliabilities. Notably, QR emerged as the subtest with the highest reliability, a distinction previously held by AR for several years.

Table 38. Cognitive Raw Score Test Statistics

Subtest	Form	Mean	*SD*	Min	Max	Alpha	*SEM*
VR (40 items)	Form 1	22.87	5.73	11	36	0.73	2.98
	Form 2	22.05	5.96	2	39	0.76	2.92
	Form 3	22.63	6.19	2	40	0.78	2.9
	Form 4	21.96	5.91	2	39	0.75	2.96
	Form 5	22.37	5.99	3	40	0.76	2.93
DM (26 items; 34 score points)	Form 1	18.87	5.52	6	31	0.74	2.81
	Form 2	18.76	5.87	2	34	0.77	2.82
	Form 3	18.94	5.86	0	34	0.78	2.75
	Form 4	18.49	5.57	1	34	0.75	2.79
	Form 5	18.36	5.8	2	34	0.76	2.84
QR (32 items)	Form 1	18.78	6.59	4	30	0.86	2.47
	Form 2	19.67	6.28	1	32	0.85	2.43
	Form 3	19.13	5.96	1	32	0.83	2.46
	Form 4	19.36	5.89	2	32	0.82	2.5
	Form 5	19.91	6.25	1	32	0.85	2.42
AR (50 items)	Form 1	33.15	6.81	18	47	0.79	3.12
	Form 2	32.09	7.35	4	49	0.82	3.12
	Form 3	33.13	7.61	5	50	0.84	3.04
	Form 4	32.05	7.51	6	50	0.83	3.1
	Form 5	32.74	8.37	5	50	0.87	3.02

Table 38 also shows the *SEM*. This value is the amount of measurement error associated with each subtest and form. *SEM* is calculated using the *SD* of the raw scores and Cronbach's alpha. Higher reliabilities result in lower *SEM*s.

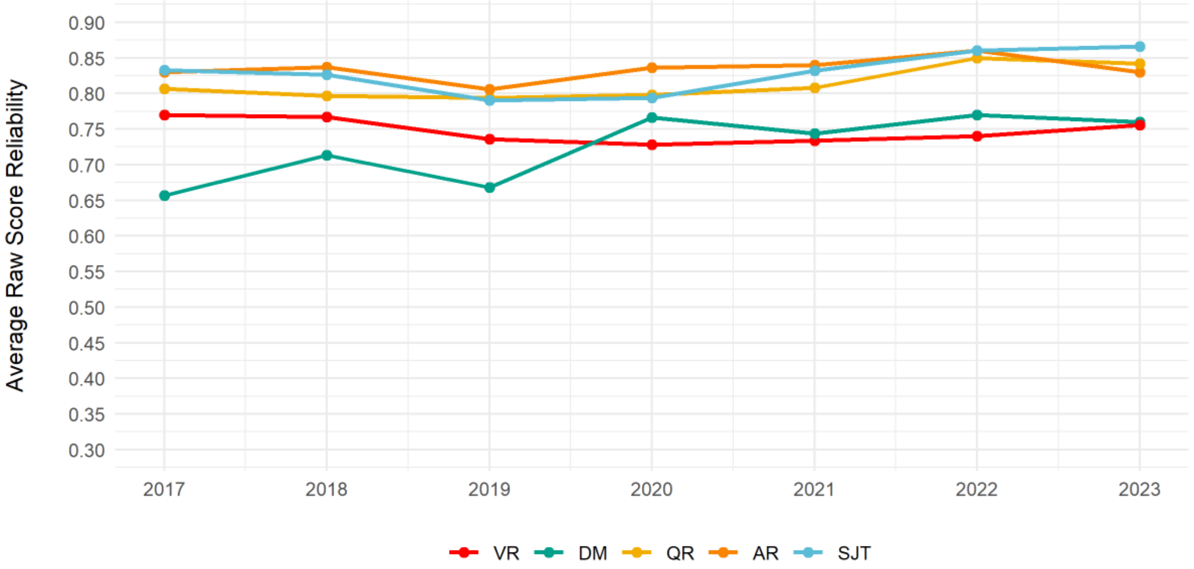
The SJT is analysed in a similar way to the cognitive sections above; however, because the maximum raw score available on the SJT can change year on year, an additional column called mean percent raw score is added (Table 39). Similar to the cognitive results, the reliability is adequately high and the *SEM* adequately low for the SJT.

Table 39. SJT Raw Score Test Statistics (252 score points)

Form	Mean	*SD*	Min	Max	Mean Percent Raw Score	Alpha	*SEM*
Form 1	199.07	21.39	123	240	78.99%	0.86	8.00
Form 2	197.65	23.53	42	240	78.43%	0.88	8.15
Form 3	197.10	21.19	70	242	78.21%	0.85	8.21
Form 4	197.86	22.14	40	242	78.52%	0.87	7.98
Form 5	196.49	23.34	56	240	77.97%	0.87	8.42

Subtest reliability has been consistent since 2017. Figure 21 shows the mean Cronbach's alpha for each subtest in each form since 2017. Note that prior to 2019, it is the mean of three forms, whereas since 2019, it is the mean of five forms. DM has become more reliable since its launch in 2017, and the reliability of VR has slightly dropped but remained consistent since 2020, with a small improvement in 2023. The reliability of both QR and the SJT has continued to improve this year.

Figure 21. Raw Score Reliability 2017–2023



Raw scores are scaled and reported as scaled scores. The summary statistics for scaled scores on each form are presented below in Table 40. Instead of alpha, the scaled score reliability is the conditional reliability at each scaled score point. Similar to the results for raw scores, the scaled score reliability is adequately high for each subtest and each form. Table 40 also includes the results for the SJT.

Table 40. Cognitive Scaled Score Test Statistics

Subtest	Form	Mean	*SD*	Min	Max	Reliability	*SEM*
VR	Form 1	597.17	72.10	450	800	0.72	38.15
	Form 2	587.83	77.58	300	900	0.75	38.79
	Form 3	596.67	80.42	300	900	0.77	38.57
	Form 4	588.44	75.85	300	900	0.74	38.68
	Form 5	590.27	77.28	300	900	0.75	38.64
DM	Form 1	620.87	85.27	420	870	0.75	42.63
	Form 2	623.74	91.79	300	900	0.77	44.02
	Form 3	624.55	93.24	300	900	0.78	43.73
	Form 4	626.76	86.42	300	900	0.75	43.21
	Form 5	617.88	89.10	300	900	0.76	43.65
QR	Form 1	637.39	90.54	430	840	0.82	38.41
	Form 2	652.52	90.40	300	900	0.81	39.40
	Form 3	643.89	84.00	300	900	0.80	37.57
	Form 4	645.35	81.46	360	900	0.79	37.33
	Form 5	656.03	91.64	300	900	0.81	39.94
AR	Form 1	655.22	79.96	500	890	0.79	36.64
	Form 2	645.53	85.00	300	900	0.79	38.95
	Form 3	660.09	93.60	300	900	0.81	40.80
	Form 4	647.50	89.11	300	900	0.81	38.84
	Form 5	656.23	105.41	300	900	0.85	40.83
Total Cognitive	Form 1	2,510.65	264.36	1,860	2,920	0.91	79.31
	Form 2	2,509.62	284.74	1,490	3,510	0.92	80.54
	Form 3	2,525.20	288.56	1,550	3,540	0.92	81.62
	Form 4	2,508.05	270.82	1,350	3,440	0.91	81.25
	Form 5	2,520.41	299.75	1,320	3,520	0.92	84.78
SJT	Form 1	608.83	72.84	350	748	0.86	27.25
	Form 2	605.57	78.30	300	748	0.88	27.12
	Form 3	602.05	72.05	300	756	0.85	27.90
	Form 4	600.52	76.25	300	754	0.87	27.49
	Form 5	599.12	77.14	300	744	0.87	27.81

7. 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.

7.1 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 greater than 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 (IRT) 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 41 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 41. Cognitive Items Passing the Quality Criteria

		VR		DM		QR		AR	
		N	%	*N*	%	*N*	%	*N*	%
Operational Scored	Pass	160	100%	104	100%	128	100%	196	98%
	Fail	0	0%	0	0%	0	0%	4	2%
	p < 10 or > 95	0	0%	0	0%	0	0%	0	0%
	pBis <= 0.1	0	0%	0	0%	0	0%	4	2%
	|b| >= 3	0	0%	0	0%	0	0%	0	0%
Pretest Unscored	Pass	235	98%	234	98%	212	97%	N/A	N/A
	Fail	5	2%	6	2%	6	3%	N/A	N/A
	p < 10 or > 95	0	0%	1	0%	0	0%	N/A	N/A
	pBis <= 0.05	4	2%	5	2%	3	1%	N/A	N/A
	|b| >= 3	1	0%	0	0%	3	1%	N/A	N/A

Consistent with previous years, only four operational items failed the analysis. Those items did not discriminate highly enough. For the pretest items, few failed in the VR, DM and QR subtests. The pretest failures were due to low discrimination in addition to items being too easy or difficult. There were no pretest items for AR this year. Figure 22 and Figure 23 show that the pretest pass rate has been consistent, with excellent pass rates for VR, DM and QR.

Figure 22. Proportion of Operational Items Failing Analysis 2017–2023

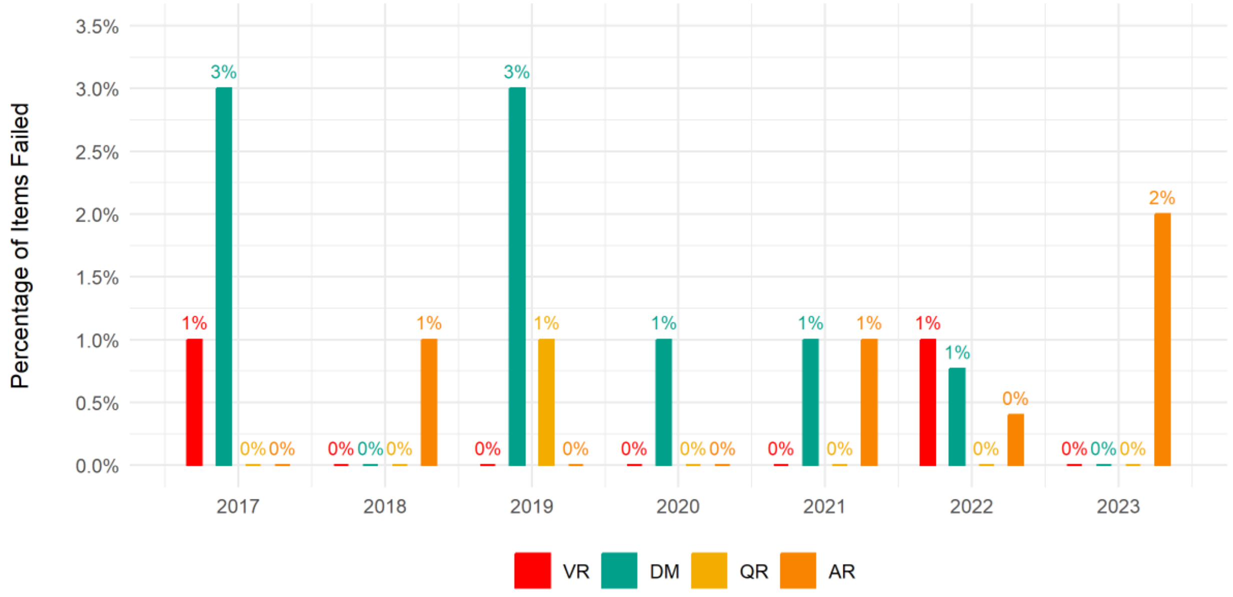


Figure 23. Proportion of Pretest Items Failing Analysis 2017–2023

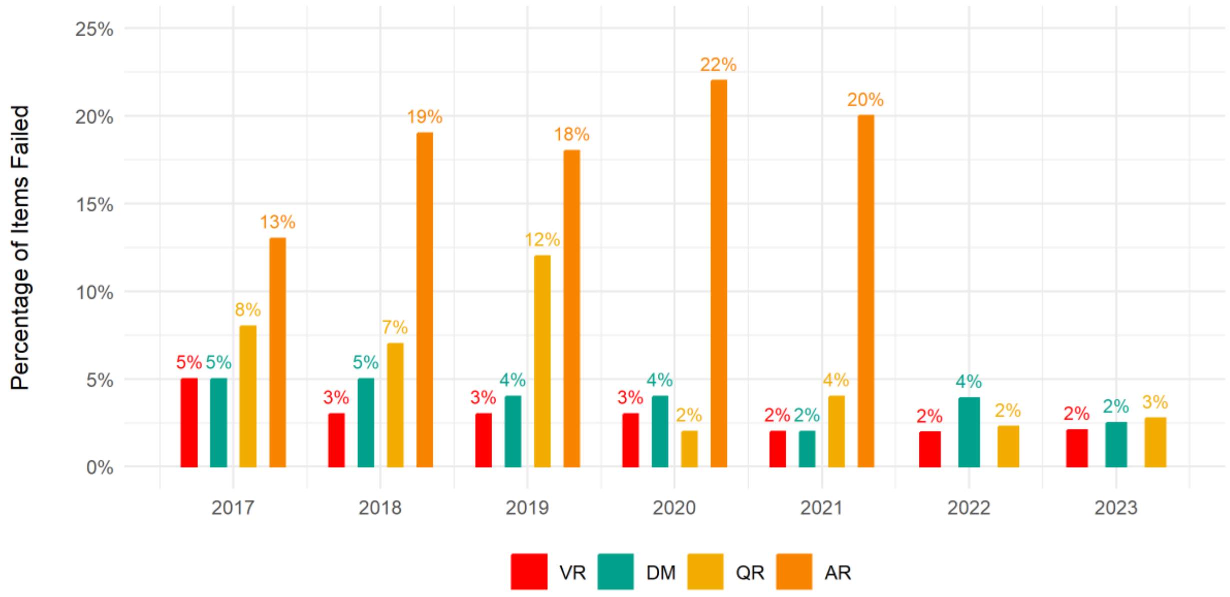


Table 42 shows a summary of the point biserial values. The maximum point biserial is 1, and higher values are better because they indicate that an item can discriminate well between strong and weak candidates. Given that the unscored items have not been tested before, it is expected that those items, on average, will discriminate less well than the scored items, and that is the case across all the cognitive subtests.

Table 42. Discrimination Summary Statistics

Scored/Unscored	Subtest	*N* Items	Mean pBis	*SD* pBis	Min pBis	Max pBis
Operational (Scored)	VR	160	0.29	0.06	0.12	0.42
	DM	104	0.37	0.08	0.14	0.57
	QR	128	0.38	0.07	0.15	0.54
	AR	200	0.32	0.09	0.04	0.54
Pretest (Unscored)	VR	240	0.25	0.08	-0.02	0.43
	DM	240	0.33	0.12	-0.06	0.60
	QR	218	0.29	0.11	-0.02	0.54
	AR	N/A	N/A	N/A	N/A	N/A

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 24. Despite a small drop in QR and AR in 2023, the operational items appear to have become slightly more discriminating over time for all subtests. This is an indication that the quality of the subtests has improved over time.

Figure 24. Point biserial 2017–2023

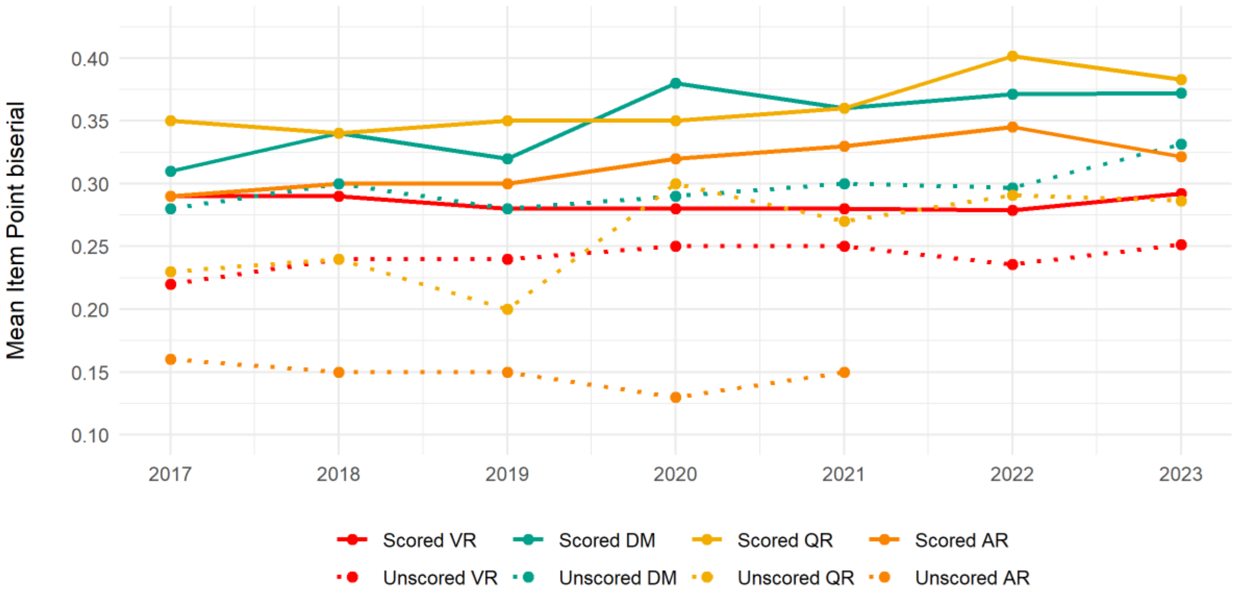


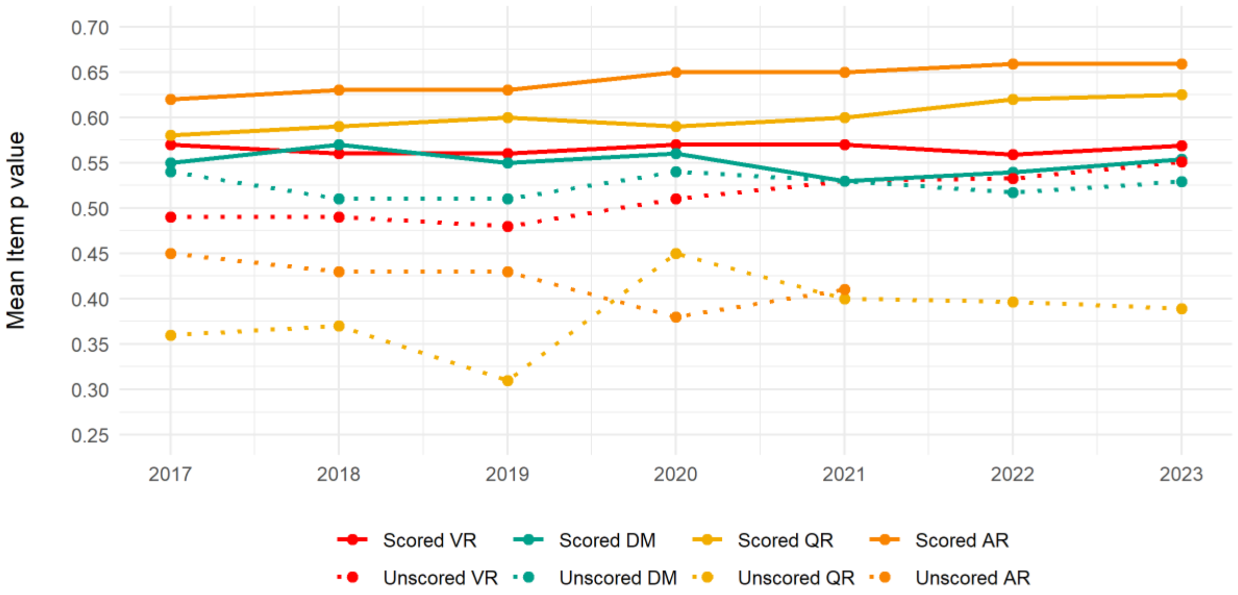
Table 43 shows the summary *of p* values for the cognitive subtests. *p* values reflect the proportion of candidates who answered an item correctly, so higher values indicate easier items, and lower values harder items. Of the operational items, DM items appear to have been the most difficult on average for 2023 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.

Table 43. *p* Value Summary Statistics

Scored/Unscored	Subtest	*N* Items	Mean *p*	*SD p*	Min *p*	Max *p*
Operational (Scored)	VR	160	0.57	0.13	0.19	0.84
	DM	104	0.55	0.15	0.24	0.93
	QR	128	0.62	0.14	0.34	0.86
	AR	200	0.66	0.13	0.23	0.91
Pretest (Unscored)	VR	240	0.55	0.17	0.04	0.94
	DM	240	0.53	0.19	0.12	0.96
	QR	218	0.39	0.17	0.07	0.93
	AR	N/A	N/A	N/A	N/A	N/A

Since 2017, pretesting has been successful in identifying items that are too difficult and too easy. Figure 25 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 25. *p* Value 2017–2023



The VR subtest consists of four-option multiple-choice items and three-option true/false/can't tell items.

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

Scored/Unscored	Item Type	*N* Items	Point biserial		*p* Value	
			Mean	*SD*	Mean	*SD*
Operational (Scored)	Multiple Choice	112	0.31	0.05	0.56	0.12
	True/False/Can't Tell	48	0.24	0.06	0.59	0.14
Pretest (Unscored)	Multiple Choice	136	0.28	0.07	0.55	0.16
	True/False/Can't Tell	104	0.22	0.09	0.56	0.19

The DM subtest contains multiple-choice items, scored out of one, and drag-and-drop items, which are scored out of two. The drag-and-drop items are more difficult than the multiple-choice items and they discriminate better, as shown in Table 45.

Table 45. DM Response Type Point biserial and *p* Value

Scored/Unscored	Response Type	*N* Items	Point biserial		*p* Value	
			Mean	*SD*	Mean	*SD*
Operational (Scored)	Drag and Drop	32	0.44	0.08	0.55	0.13
	Multiple Choice	72	0.34	0.06	0.56	0.16
	Drag and Drop	63	0.44	0.10	0.47	0.18

Scored/Unscored	Response Type	*N* Items	Point biserial		*p* Value	
			Mean	*SD*	Mean	*SD*
Pretest (Unscored)	Multiple Choice	177	0.29	0.11	0.55	0.18

In addition to different response types, the DM subtest also contains different item types. Among the drag-and-drop items, interpreting information items are more difficult than syllogism items but the latter discriminate slightly better than the former, as presented in Table 46. For the multiple-choice items, the items on statistical reasoning and Venn diagrams are the most discriminating. Logical Puzzles were found to be the most difficult item type in DM, while Syllogisms were found to be the easiest.

Table 46. DM Response and Item Type Point biserial and *p* Value

Scored/ Unscored	Response Type	Item Type	*N* Items	Point biserial		*p* Value	
				Mean	*SD*	Mean	*SD*
Operational (Scored)	Drag and Drop	Information Interpretation	16	0.40	0.07	0.48	0.09
		Syllogisms	16	0.49	0.06	0.61	0.14
	Multiple Choice	Logical Puzzles	16	0.32	0.05	0.43	0.08
		Statistical Reasoning	16	0.38	0.05	0.56	0.10
		Assumptions Recognition	16	0.29	0.07	0.64	0.20
		Venn Diagrams	24	0.36	0.05	0.58	0.16
Pretest (Unscored)	Drag and Drop	Information Interpretation	30	0.39	0.10	0.40	0.16
		Syllogisms	33	0.48	0.08	0.53	0.18
	Multiple Choice	Logical Puzzles	32	0.30	0.08	0.55	0.20
		Statistical Reasoning	25	0.33	0.09	0.46	0.13
		Assumptions Recognition	50	0.23	0.12	0.55	0.17
		Venn Diagrams	70	0.32	0.10	0.58	0.20

The QR subtest has item sets and standalone items. Each item set contains four items. As with the pretest pool as a whole, the pretest items discriminate less well on average than the ones that have already been pretested prior to appearing in the 2023 exam, as shown in Table 47.

Table 47. QR Type Point biserial and *p* Value

Scored/Unscored	Item Type	*N* Items	Point biserial		*p* Value	
			Mean	*SD*	Mean	*SD*
Operational (Scored)	Item Set	112	0.38	0.07	0.61	0.14
	Standalone	16	0.37	0.06	0.71	0.14
Pretest (Unscored)	Item Set	198	0.28	0.10	0.37	0.16
	Standalone	20	0.35	0.10	0.56	0.19

The AR subtest consists of four different types. Table 48 below shows that the discrimination of all four item types is similarly strong across the operational items.

Table 48. AR Type Point biserial and *p* Value

Scored/Unscored	Item Type	*N* Items	Point biserial		*p* Value	
			Mean	*SD*	Mean	*SD*
Operational (Scored)	Type 1	160	0.32	0.09	0.65	0.13
	Type 2	8	0.29	0.04	0.67	0.23
	Type 3	12	0.26	0.05	0.74	0.16
	Type 4	20	0.36	0.08	0.70	0.11

7.1.1 Item Analysis for SEN

An additional analysis was performed this year to examine whether the items perform differently for exams with accommodations. Overall, the item performances did not show substantial differences between the two set of analyses, with all of the differences being within a third of an *SD* and most of them being within a tenth of an *SD*, as presented in Table 49. The item analysis performed using the UCATSEN sample consistently showed a higher *p* value, which is consistent with the higher performance of the UCATSEN candidates when compared to the UCAT candidates, as reported in the previous section. Most of the average IRT *b* values across the two sets of analyses are identical and the largest difference is less than a tenth of an *SD*, showing that the items present similar item difficulties to candidates in both exam codes after considering their ability level.

Table 49. Item Analysis of UCAT and UCATSEN

Scored/Unscored	Subtest	Statistics	UCAT		UCATSEN	
			Mean	SD	Mean	SD
Operational (Scored)	VR	*p* Value	0.57	0.13	0.61	0.13
		Point biserial	0.29	0.06	0.29	0.08
		IRT *b*	-0.21	0.61	-0.19	0.62
	DM	Facility	0.72	0.32	0.78	0.34
		Point biserial	0.37	0.08	0.37	0.08
		IRT *b*	0.23	0.72	0.21	0.77
	QR	*p* Value	0.62	0.14	0.66	0.15
		Point biserial	0.38	0.07	0.37	0.07
		IRT *b*	-0.26	0.71	-0.27	0.78
	AR	*p* Value	0.66	0.13	0.70	0.14
		Point biserial	0.32	0.09	0.31	0.11
		IRT *b*	0.15	0.73	0.15	0.75
Pretest (Unscored)	VR	*p* Value	0.55	0.17	0.58	0.20
		Point biserial	0.25	0.09	0.24	0.22
		IRT *b*	-0.14	0.90	-0.09	1.08
	DM	Facility	0.65	0.30	0.68	0.34
		Point biserial	0.33	0.13	0.33	0.25

Scored/Unscored	Subtest	Statistics	UCAT		UCATSEN	
			Mean	SD	Mean	SD
		IRT *b*	0.31	0.95	0.44	1.07
	QR	*p* Value	0.39	0.17	0.41	0.20
		Point biserial	0.28	0.11	0.31	0.20
		IRT *b*	0.98	0.95	1.05	1.16

7.1.2 Comparison of UCAT Item Bank Statistics with UCAT ANZ

The following section is an updated version of the same comparison made in this year's UCAT ANZ technical report with updated item statistics from UCAT 2023. This section presents the test items performance across the UK and ANZ population of the 2023 cohort. It should be noted that both the *p* value and point biserial are classical statistics and are therefore dependent upon the performance of the group on which the test was administered. The IRT difficulty, on the other hand, is anchored back to a common benchmark, so these values are comparable across windows.

Table 50 compares the summary statistics for the operational item analysis for the UCAT 2023 to the UCAT ANZ 2023 values. Across all the subtests, the point biserial summary statistics were similar, with the results from the ANZ population showing slightly higher values, indicating that all operational items discriminated as strongly as expected for the UCAT ANZ population. In terms of the *p* value, which is sample-dependant, the UCAT ANZ population had higher (i.e. easier) average values across subtests. The IRT difficulty, on the other hand, is on a common scale. Table 50 shows that for all subtests, the 2023 UCAT and UCAT ANZ had very similar mean IRT difficulty values, indicating a comparable level of difficulty for both populations.

Table 50. Comparison of Operational Item Statistics: UCAT & UCAT ANZ 2023

Subtest	Item Statistics	*N* Items	UCAT 2023		UCAT ANZ 2023	
			Mean	*SD*	Mean	*SD*
VR	*p* Value	160	0.57	0.13	0.60	0.13
	Point biserial	160	0.29	0.06	0.30	0.06
	IRT Difficulty	160	-0.21	0.61	-0.19	0.61
DM	Facility	104	0.55	0.15	0.58	0.15
	Point biserial	104	0.37	0.08	0.39	0.09
	IRT Difficulty	104	0.23	0.72	0.20	0.72
QR	*p* Value	128	0.62	0.14	0.65	0.13
	Point biserial	128	0.38	0.07	0.41	0.07
	IRT Difficulty	128	-0.26	0.71	-0.26	0.72
AR	*p* Value	200	0.66	0.13	0.67	0.13
	Point biserial	200	0.32	0.09	0.35	0.10
	IRT Difficulty	200	0.15	0.73	0.17	0.73

In addition, during the standard UCAT and UCAT ANZ item analysis, any item that shows an item drift more extreme than +/-0.5 is removed from the anchor and re-calibrated as the item difficulty is considered to have changed significantly. This can give an indication of whether the relative difficulty of the items for the UCAT ANZ population is comparable to that for the UCAT population.

Table 51 summarises the number of items showing drift in the UCAT since 2017 compared to the UCAT ANZ since 2019. Compared to the UCAT 2023, the number of drift items are slightly higher for the UCAT ANZ 2023. These items were reviewed by the Content Team and there was no clear explanation for the differences in terms of the cultural sensitivity of the items.

Table 51. Number of Operational Items Showing Drift in UCAT vs UCAT ANZ

Subtest	UCAT							UCAT ANZ				
	2017	2018	2019	2020	2021	2022	2023	2019	2020	2021	2022	2023
VR	2 (2%)	3 (3%)	6 (3%)	4 (2%)	4 (2%)	5 (3%)	6 (4%)	12 (10%)	13 (6%)	13 (6%)	8 (4%)	9 (6%)
DM	11 (14%)	6 (8%)	17 (13%)	37 (28%)	12 (9%)	7 (5%)	3 (3%)	7 (9%)	47 (36%)	11 (8%)	9 (7%)	8 (8%)
QR	2 (2%)	0 (0%)	1 (1%)	0 (0%)	2 (1%)	6 (4%)	2 (2%)	3 (3%)	2 (1%)	4 (2%)	5 (3%)	4 (3%)
AR	7 (5%)	5 (3%)	21 (8%)	25 (10%)	40 (16%)	19 (8%)	5 (3%)	22 (15%)	24 (10%)	37 (15%)	13 (5%)	7 (4%)

At present, it is recommended that the degree of drift is monitored in 2024. We would not recommend taking any action to create a separate item bank for the UCAT ANZ at this time.

7.2 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 52. 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 52. Candidate Removal Summary for SJT Item Analysis

Statistic	Criteria	Number of Candidates Removed
1. *Z* score of the scaled score	*Z* score < -4.189	0
2. High number of missing responses	> 1 blank response on operational items	840
3. Low completion time	Drop in score based on response time	0

Statistic	Criteria	Number of Candidates Removed
4. Form 1 Candidates	Candidates who attempted Form 1	46

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

In 2023, there are discussions in adjusting the SJT item quality criteria to align them with the criteria used for the cognitive items. The changed criteria are expected to be slightly more lenient than the current criteria which will result in more operational and pretest items being deemed successful and support with the continued development and improvements to the item bank. As these changes are still under discussion and the following results are based on the existing quality criteria.

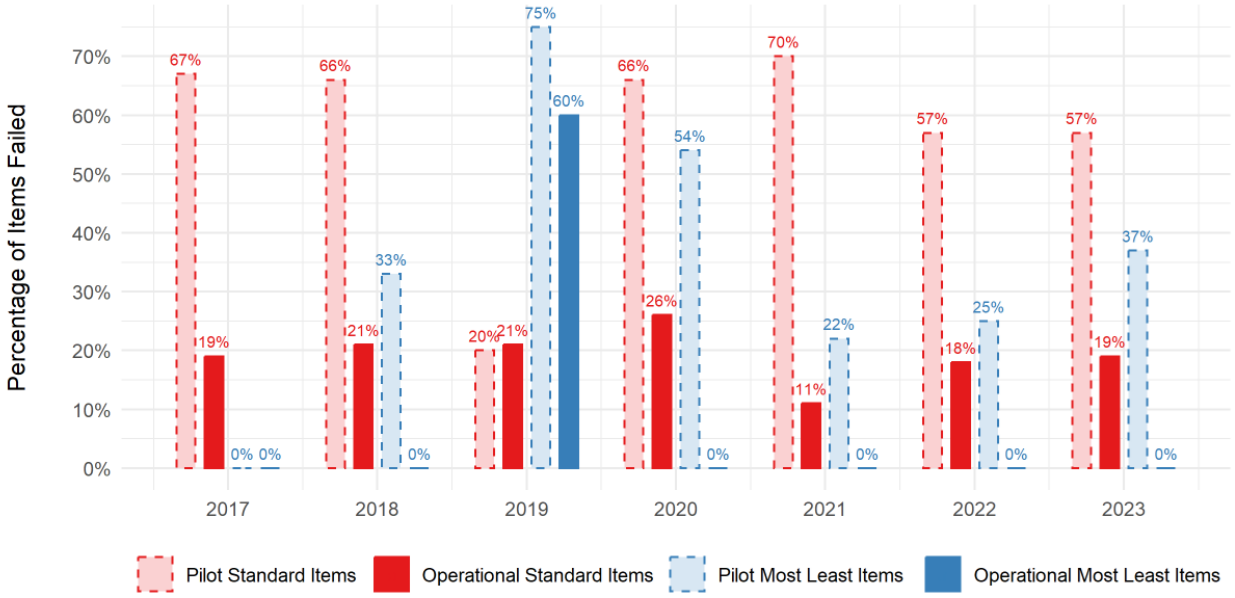
Table 53 shows the number of items that met and did not meet the quality criteria. The most/least item type was more successful than the standard items, with all operational items and 63% of the pretest items meeting the criteria.

Table 53. SJT Item Quality Criteria

	Item Type	Statistical Criteria Met/Not Met	All		Appropriateness		Direct Speech		Importance	
			N	%	*N*	%	*N*	%	*N*	%
Operational	Rating Items	Met	154	81%	61	78%	23	70%	70	90%
		Not met	35	19%	17	22%	10	30%	8	10%
	Most/Least Items	Met	6	100%						
		Not met	0	0%						
Pretest	Rating Items	Met	139	43%	99	42%	15	41%	25	47%
		Not met	185	57%	135	58%	22	59%	28	53%
	Most/Least Items	Met	17	63%						
		Not met	10	37%						

The proportion of items meeting the quality criteria is fairly consistent with previous years. Figure 26 shows that the proportion of operational standard items meeting the criteria is consistent between 2022 and 2023. The number of pretest most/least items not meeting the criteria increased slightly from 25% in 2022 to 37% in 2023. In 2023, the percentage of standard rating items that met the criteria was 43% and it is the same as the percentage of the pretest dichotomous items meeting the criteria in 2022. However, it should be noted that a proportion of the dichotomous items pretested in 2022 were adapted from already successful items in the item bank and therefore a higher pass rate is expected. A similar passing rate in newly developed items in 2023 showed an improvement in pretest item quality.

Figure 26. Proportion of SJT Items Failing Analysis 2017–2023



The summary of all operational SJT items is shown below in Table 54.

Table 54. Operational SJT Item Analysis Summary

	Mean	*SD*	Min	Max
Item Mean	3.06	1.05	0.41	7.42
Item *SD*	1.00	0.31	0.27	2.08
Item Partial Correlation	0.29	0.13	-0.05	0.55
Item Total Facility	0.77	0.16	0.14	0.99

Since 2017, the item mean score and facility has tended to increase, as illustrated in Figure 27, indicating that items are becoming somewhat easier. Figure 28 shows an increase in item partial correlation, which indicates that despite the test being relatively easy, it has progressively improved in consistently measuring the same ability, and the items are getting better overall at discriminating among strong and weak candidates. A better discrimination between candidates implies that the test results could be considered as being more reliable in distinguishing stronger and weaker candidates. In other words, improvement is seen in the item quality. However, a relatively high facility could imply that the test might be too easy to distinguish between strong and very strong candidates. Even though high facility items provide face validity to the SJT, in future test development, harder items will also be developed to minimise the upward trend of item facility.

Figure 27. Average Item Facility of Operational SJT Items 2017–2023

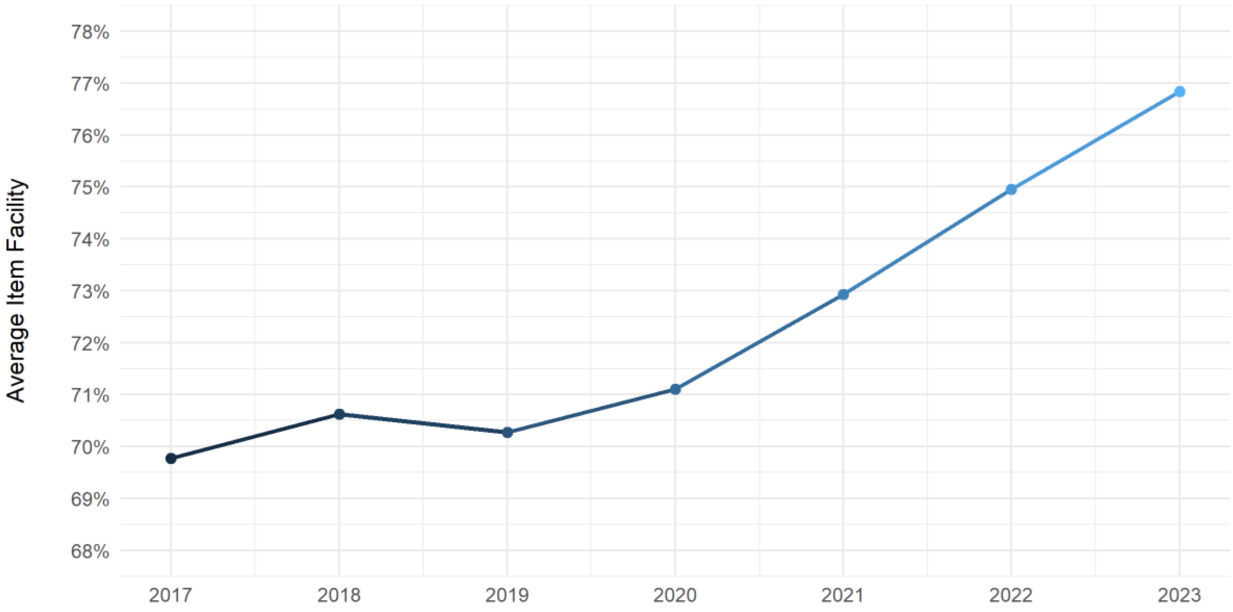


Figure 28. Average Item Partial Correlation of Operational SJT Items 2017–2023

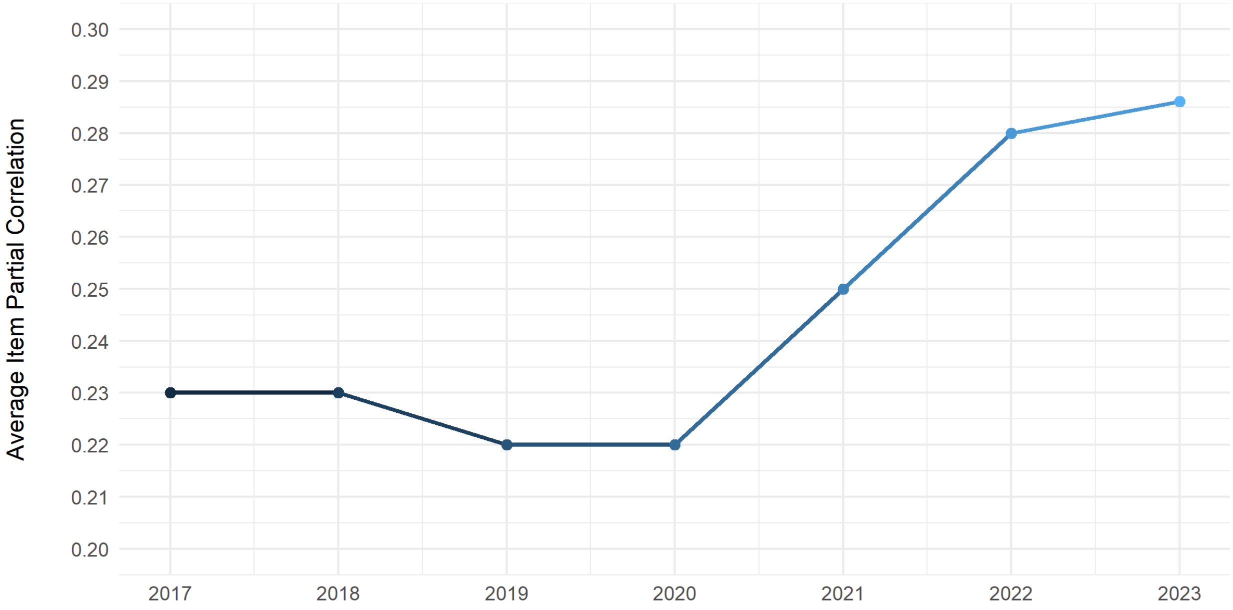


Table 55 shows the summary statistics for the SJT pretest items. While the Most/Least items showed a slightly higher discriminating ability than the standard rating items, the average item total facility is relatively high.

Table 55. SJT Pretest Item Summary Statistics

	Statistic	Item Mean	Item *SD*	Item Partial	Item Total Facility
Rating Items	Mean	2.89	0.89	0.15	0.76
	SD	0.89	0.31	0.14	0.19
	Min	0.57	0.15	-0.17	0.19
	Max	3.98	1.54	0.48	0.99
Most/Least	Mean	7.19	1.29	0.17	0.90
	SD	0.44	0.44	0.06	0.05
	Min	5.86	0.58	0.06	0.73
	Max	7.89	2.79	0.28	0.99

7.3 Differential Item Functioning (DIF)

7.3.1 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.

7.3.2 Method of DIF Detection

For the 2023 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.

7.3.3 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.

7.3.4 DIF Results

The DIF results are reported below for each demographic group. Table 56 shows DIF in relation to gender. One operational AR item was found to exhibit Category C DIF favouring Male over Female.

Table 56. Gender DIF

Group	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	160	100%	101	97%	125	98%	199	100%	193	98%
	B	0	0%	3	3%	3	2%	0	0%	3	2%
	C	0	0%	0	0%	0	0%	1	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	A	240	100%	239	100%	217	100%	N/A	N/A	336	95%
	B	0	0%	1	0%	1	0%	N/A	N/A	15	4%
	C	0	0%	0	0%	0	0%	N/A	N/A	0	0%
	NA	0	0%	0	0%	0	0%	N/A	N/A	0	0%

In contrast to previous years, the age comparison has been changed to increase the number of items where a comparison can be made. Since 2022, the comparison is between less than 20 and greater than 25 in contrast to less than 20 and greater than 35 in previous years (Table 57). One operational VR item was found to exhibit Category C DIF favouring younger candidates. Three Category C items were identified in DM, with one item favouring older candidates and two favouring younger candidates. Two AR items showed Category C DIF; both favoured younger candidates over older candidates. These items will be reviewed by the Content Team and removed from the bank.

Table 57. Age DIF

Group	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	156	98%	88	85%	127	99%	195	98%	196	100%
	B	3	2%	13	12%	1	1%	3	2%	0	0%
	C	1	1%	3	3%	0	0%	2	1%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	A	0	0%	0	0%	0	0%	N/A	N/A	341	97%
	B	0	0%	0	0%	0	0%	N/A	N/A	10	3%
	C	0	0%	0	0%	0	0%	N/A	N/A	0	0%
	NA	240	100%	240	100%	218	100%	N/A	N/A	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. Note that the options on the ethnicity question have changed since 2022 and the “UK - Chinese” category is no longer separate. In addition, a comparison between White and Non-White was included.

Table 58 shows there were six instances of Category C DIF identified in the ethnicity comparisons. Of these, one operational AR item favoured White candidates over Black candidates; one pretest QR item favoured White candidates over both Asian and non-White candidates; and one SJT pretest item showed a reverse pattern of favouring Asian and non-White candidates over White candidates.

Table 58. Ethnicity DIF

Type	Group	Code	VR		DM		QR		AR		SJT	
			N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	White/ Black	A	156	98%	100	96%	127	99%	197	98%	179	91%
		B	4	2%	4	4%	1	1%	2	1%	17	9%
		C	0	0%	0	0%	0	0%	1	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
	White/ Asian	A	159	99%	102	98%	128	100%	200	100%	179	91%
		B	1	1%	2	2%	0	0%	0	0%	17	9%
		C	0	0%	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%	0	0%
	White/ Mixed	A	159	99%	103	99%	128	100%	200	100%	196	100%
		B	1	1%	1	1%	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%
	White/ Non-White	A	160	100%	102	98%	128	100%	200	100%	187	95%
		B	0	0%	2	2%	0	0%	0	0%	9	5%
		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	42	18%	7	3%	140	64%	N/A	N/A	79	23%
		B	0	0%	1	0%	0	0%	N/A	N/A	9	3%
		C	0	0%	0	0%	0	0%	N/A	N/A	0	0%
		NA	198	82%	232	97%	78	36%	N/A	N/A	0	0%
	White/ Asian	A	240	100%	215	90%	217	100%	N/A	N/A	322	92%
		B	0	0%	0	0%	0	0%	N/A	N/A	15	4%
		C	0	0%	0	0%	1	0%	N/A	N/A	1	0%
		NA	0	0%	25	10%	0	0%	N/A	N/A	0	0%
	White/ Mixed	A	0	0%	0	0%	0	0%	N/A	N/A	74	21%
		B	0	0%	0	0%	0	0%	N/A	N/A	0	0%
		C	0	0%	0	0%	0	0%	N/A	N/A	0	0%
		NA	240	100%	240	100%	218	100%	N/A	N/A	0	0%
	White/ Non-White	A	239	100%	240	100%	216	99%	N/A	N/A	337	96%
		B	0	0%	0	0%	0	0%	N/A	N/A	13	4%
		C	1	0%	0	0%	2	1%	N/A	N/A	1	0%
		NA	0	0%	0	0%	0	0%	N/A	N/A	0	0%

Since 2022, a comparison between SEC1 and non-SEC-1 has also been included to allow more comparisons to be made. One pretest QR item was categorised as DIF category C, favouring SEC-1 candidates over non-SEC-1 candidates. Four category C DIF items were identified for SJT pretest items, all favouring SEC-1 candidates over non-SEC-1, SEC-2, SEC-3, and SEC-4 respectively.

Table 59. SEC DIF

Type	Group	Code	VR		DM		QR		AR		SJT	
			N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	SEC 1/2	A	160	100%	104	100%	128	100%	200	100%	196	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	160	100%	104	100%	128	100%	200	100%	196	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	160	100%	104	100%	128	100%	200	100%	196	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	160	100%	104	100%	127	99%	200	100%	196	100%
		B	0	0%	0	0%	1	1%	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/(2-5)	A	160	100%	104	100%	128	100%	200	100%	196	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%
Pretest	SEC 1/2	A	0	0%	0	0%	0	0%	N/A	N/A	216	62%
		B	0	0%	0	0%	0	0%	N/A	N/A	5	1%
		C	0	0%	0	0%	0	0%	N/A	N/A	1	0%
		NA	240	100%	240	100%	218	100%	N/A	N/A	0	0%
	SEC 1/3	A	138	57%	40	17%	167	77%	N/A	N/A	265	75%
		B	0	0%	0	0%	0	0%	N/A	N/A	6	2%
		C	0	0%	0	0%	0	0%	N/A	N/A	1	0%
		NA	102	42%	200	83%	51	23%	N/A	N/A	0	0%
	SEC 1/4	A	0	0%	0	0%	0	0%	N/A	N/A	237	68%
		B	0	0%	0	0%	0	0%	N/A	N/A	1	0%
		C	0	0%	0	0%	0	0%	N/A	N/A	1	0%
		NA	240	100%	240	100%	218	100%	N/A	N/A	0	0%
	SEC 1/5	A	1	0%	0	0%	8	4%	N/A	N/A	246	70%
		B	0	0%	0	0%	0	0%	N/A	N/A	3	1%
		C	0	0%	0	0%	0	0%	N/A	N/A	0	0%
		NA	239	100%	240	100%	210	96%	N/A	N/A	0	0%
	SEC 1/(2-5)	A	240	100%	232	97%	217	100%	N/A	N/A	339	97%
		B	0	0%	0	0%	0	0%	N/A	N/A	6	2%
		C	0	0%	0	0%	1	0%	N/A	N/A	1	0%
		NA	0	0%	8	3%	0	0%	N/A	N/A	0	0%

As Table 60 illustrates, there was one Category C DIF item detected in the comparison between candidates who had an honours degree or above and those who did not. This item was a QR pretest item and favoured candidates with a degree education over those without. There were high candidate volumes across the board, meaning comparisons could be made for all subtests.

Table 60. Honours Degree DIF

Type	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	160	100%	102	98%	127	99%	200	100%	196	100%
	B	0	0%	2	2%	1	1%	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	A	240	100%	190	79%	217	100%	N/A	N/A	344	98%
	B	0	0%	0	0%	0	0%	N/A	N/A	7	2%
	C	0	0%	0	0%	1	0%	N/A	N/A	0	0%
	NA	0	0%	50	21%	0	0%	N/A	N/A	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 61 shows, two pretest items were flagged as having Category C DIF. The pretest DM item was found to favour native English speakers over non-native English speakers while the pretest SJT item was found to favour non-native English speakers over native English speakers.

Table 61. English as First Language DIF

Group	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	160	100%	104	100%	128	100%	200	100%	195	99%
	B	0	0%	0	0%	0	0%	0	0%	1	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	240	100%	238	99%	218	100%	N/A	N/A	330	94%
	B	0	0%	1	0%	0	0%	N/A	N/A	20	6%
	C	0	0%	1	0%	0	0%	N/A	N/A	1	0%
	NA	0	0%	0	0%	0	0%	N/A	N/A	0	0%

Four Category C DIF items were identified in the comparison of candidates who reported the UK as their residence with those who reported the UK as not being their residence (as presented in Table 62). A pretest VR item and a pretest SJT item were found to favour UK residents over non-UK residents; and a pretest VR item and a pretest QR item were found to show the reverse pattern.

Table 62. Residency DIF

Group	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
Operational	A	159	99%	103	99%	124	97%	200	100%	189	96%
	B	1	1%	1	1%	4	3%	0	0%	7	4%
	C	0	0%	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	A	238	99%	240	100%	217	100%	N/A	N/A	321	91%
	B	0	0%	0	0%	0	0%	N/A	N/A	29	8%

Group	Code	VR		DM		QR		AR		SJT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
	C	2	1%	0	0%	1	0%	N/A	N/A	1	0%
	NA	0	0%	0	0%	0	0%	N/A	N/A	0	0%

Very few candidates took the online version of the UCAT (31 candidates; see Section 4.4), so comparison was not possible.

In conclusion, 24 Category C DIF items were identified, with 8 operational items and 16 pretest items. This is a higher number compared to 2022, where only 10 Category C DIF items were identified. This increase may be partially attributed to a slight increase in number of candidates from ethnic minority groups, particularly UK - Asian and non-UK candidates. The greater diversity among candidates might have contributed to more varied responses to the items, aiding in the detection of item bias. These items have been removed from the item bank to ensure they are not used in future tests, and additional efforts will be made to review these items to reduce potential bias in future item development.

8. Summary

In 2023, the subtests underwent a rescaling process to reduce the differences in average scaled scores between them. This rescaling achieved its purpose, with the differences in averaged scaled scores across subtests being notably reduced.

The scores in the 2023 administration of the UCAT were broadly in line with scores in previous years. The proportion of candidates taking the SEN version remained unchanged, and the demographic composition of the test-takers stayed mostly the same, except for a continued decline in UK - White candidates and an increase in UK - Asian and non-UK candidates.

Candidates taking a SEN version of the exam continue to score better than candidates taking the non-SEN version, and demographic trends in scores and candidate volumes were consistent with previous years' administrations of the exam. Higher scores continue to be associated with candidates who are resident in the UK, have White ethnicity, are in SEC 1, and speak English as a first language. Certain scoring patterns by demographic also persist in the 2023 version of the exam. Male candidates outperformed female candidates on the cognitive sections and vice versa on the SJT.

In terms of test quality, the test forms were reliable, with appropriately low measurement error, and individual items performed well, with very few operational cognitive items needing to be retired. More SJT items did not meet the required criteria than the operational items which is consistent with performance in previous years. However, SJT criteria is currently under review so that it is aligned with that of the cognitive tests which will result in more items meeting the criteria. There are a relatively higher number of Category C DIF items identified this year.

8.1 Recommendations

The outcome of the UCAT 2023 analysis identifies certain small operational changes that have improved 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. AR and QR, as the highest scoring subtests, have a greater influence on the total score than VR, which is the lowest scoring subtest. In 2023, the subtests were slightly rescaled to bring them closer to an average score of 600, which was found to be effective. Pearson VUE recommend continuing the rescaling year-on-year until a more balanced scaled scores distribution is achieved.

In 2022, adjustments were implemented to reduce the speededness of the subtests and these modifications were continued in 2023, resulting in a level of speededness comparable to 2022 but less than prior years. The analysis on speededness after excluding guessing behaviour showed that candidates generally managed to address most items in each subtest. Nonetheless, the issue of speededness is still potentially a concern that will be continue to be monitored. Currently Pearson VUE constructs test forms with a constraint on the historic average response time for all of the subtests to minimise this. This constraint means item times are balanced in the form-building process so that all of the test forms have a similar number of time-consuming items and a similar number of less time-consuming items. As a result, the average time to answer all items on the subtest should be kept to a reasonable level, and this will continue to be monitored. No changes to the test timings are recommended for 2024 at present.