Pearson

University Clinical Aptitude Test (UCAT)

Technical Report

7 July 2025 to 26 September 2025

Non-Disclosure and Confidentiality Notice

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

Table of Contents

Table of Tables

Table of Figures

Executive Summary

The University Clinical Aptitude Test (UCAT) was administered in 2025 between 7 July 2025 and 26 September 2025. There were notable changes in 2025 in candidate volumes, exam structure, and scoring approaches. The exam underwent significant restructuring, including the removal of the Abstract Reasoning (AR) subtest, additional time and items allocated to Decision Making (DM), and extra time for Verbal Reasoning (VR) and Quantitative Reasoning (QR). To account for these changes, VR and QR subtest scores were downscaled by 10 points.

A total of 41,354 exams were administered, marking a close to 10% increase from the previous year. The vast majority of candidates took the exam in test centres, with online delivery accounting for less than 1% and thus not suitable for comparative analysis. Special educational needs (SEN) accommodations continued to be offered, with seven UCAT versions available. The UCATSEN version was most frequently used, and, as in prior years, SEN candidates generally outperformed those taking the standard version.

The 2025 UCAT comprised five test forms, each carefully balanced to ensure fairness and consistency across all candidates. The distribution of candidate performance within the SJT bands in 2025 closely aligned with the target distribution established for the assessment. This alignment is largely attributed to adjustments made in the SJT band-setting methodology for this year.

Reliability was consistently high across all test forms, with standard errors of measurement remaining low and stable, demonstrating effective calibration of form difficulty and candidate performance. The cognitive subtests retained their speeded nature, as most candidates used nearly all allotted time. However, with additional time granted for the QR and VR sections, the degree of speededness in both subtests was reduced. Speededness was lowest among SEN candidates who received extended time, while the SJT subtest continued to exhibit the least speededness overall.

The composition of candidates in 2025 remains largely consistent with that observed in the previous year. However, a significant change this year is the record number of candidates sitting the exam. This continued expansion of the exam is demonstrated by a notable increase in candidate volume, with 2025 registering nearly 10% more candidates compared to 2024. Furthermore, when compared to candidate volume from 2017, there is an increase of more than 65%, underscoring the substantial growth in candidate numbers over recent years.

The item analysis for this year demonstrates continued improvement in the quality of items within the bank. Notably, all operational items in the cognitive subtests met the required statistical criteria, which reflects a gradual enhancement in item quality over time. This achievement highlights the rigorous development and review processes in place for operational items. Consistent

with trends observed in previous years, the passing rates for pretest items within the cognitive subtests remained exceptionally high, suggesting that the pretest item pool continues to meet the expected standards for candidate performance and item validity. Similarly, the passing rates for the SJT subtest were largely in line with previous years' results.

In conclusion, the results of the 2025 UCAT administration were broadly consistent with those of previous years. Test forms demonstrated high reliability, low measurement error, and balanced difficulty across forms. Despite changes to exam design and the increased candidate volume, the UCAT continues to serve as a reliable tool for identifying individuals with the potential to develop into effective clinicians.

1. Introduction

The purpose of the UCAT is to help select and/or identify more accurately those individuals with the innate ability to develop the professional skills and competencies required to be good medical and dental students. It is not an exam that measures student achievement, and therefore it does not contain any curriculum or science content.

This report covers the 2025 UCAT that was delivered from 7 July 2025 to 26 September 2025. As outlined in Section 3, the exam consisted of four subtests that each contained between 35 and 69 items. The design of the exam underwent major restructuring this year, with the AR subtest removed, additional time and items allocated to DM, and additional time given to VR and QR. The subtest scores for VR and QR were downscaled by 10 points to account for the additional allocated time.

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 characteristics, 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.

2. Exam Design 2025

The 2025 UCAT consisted of five balanced test forms, each with four subtests. Each subtest included scored and unscored items as shown in Table 1 below.

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	22 minutes allowed on items and 1.5 minutes for instruction
DM	1 testlet of 31 items	4 items	35	37 minutes allowed on items and 1.5 minutes for instruction
QR	8 testlets of 4 items	1 testlet of 4 items	36	26 minutes allowed on items and 2 minutes for instruction
AR	Removed in 2025			
SJT	20 testlets of 1 to 4 items	2 testlets of 1 to 5 items	69	26 minutes allowed on items and 1.5 minutes for instruction

Candidates were given just under 2 hours to answer a total of 184 items from the four subtests. There were six groups of candidates who took a SEN version of the exam, and thus had extra time allowances in 2025. The timing and scoring of the SEN exams are explored in detail in Section 4.2.

Over the past decade, candidate scores on the AR subtest increased while response times decreased, illustrating its coachable and mechanical nature and raising concerns about its content validity. Additionally, studies found the AR subtest had weaker predictive and incremental validity compared to other cognitive subtests (Bala, Pedder, Sam, & Brown, 2022; Paton, McManus, Cheung, Smith, & Tiffin, 2022; Greatrix, Nicholson, & Anderson, 2021; Tiffin, et al., 2016). As a result, the AR subtest was removed from the exam in 2025.

After the removal of the AR subtest from the exam, the time previously allocated to this section was redistributed among the remaining subtests in order to reduce the impact of exam speededness. The VR subtest was allotted an additional minute, increasing its total duration to 22 minutes in 2025. Due to this change, candidates had more time to complete the VR section, which effectively reduced its difficulty. As a result, VR scores were scaled down by 10 points to adjust for this lowered level of difficulty. Similarly, the QR subtest received one extra minute, bringing its total duration to 26 minutes. QR scores were also scaled down by 10 points for the same reason. The number of operational items for QR and VR remained the same as in previous years. In contrast, the number of operational items for DM was increased by five, making a new total of 31, and an

additional pretest item was included, resulting in four pretest items overall. To account for these changes and to ensure sufficient time for completion, six extra minutes were provided, bringing the total duration for DM to 37 minutes. No score scaling was applied to DM. The SJT subtest remained unchanged.

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 840. Universities received the cognitive subtest scaled scores plus a total score: a simple sum of the three cognitive subtest scores ranging from 900 to 2,700. 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	658–900	22%	Those in Band 1 demonstrated an excellent level of performance, showing similar judgement in most cases to the panel of experts.
Band 2	596–657	38%	Those in Band 2 demonstrated a good, solid level of performance, showing appropriate judgement frequently, with many responses matching model answers.
Band 3	502–595	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	300–501	10%	The performance of those in Band 4 was low, with judgement tending to differ substantially from ideal responses in many cases.

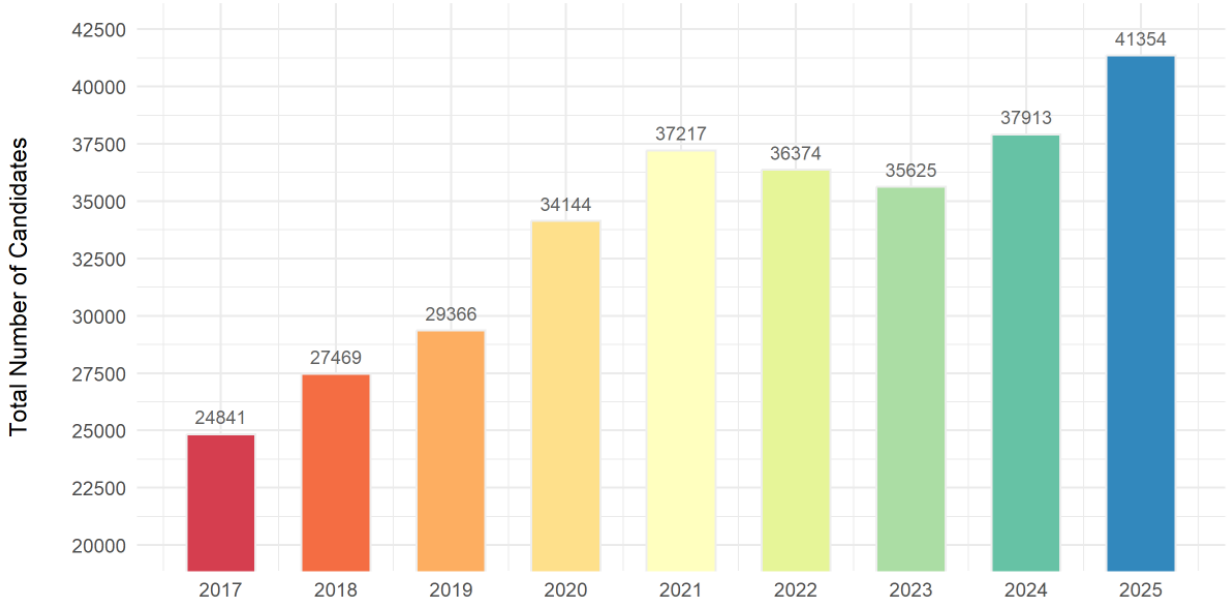
The 2025 UCAT was delivered in two modes: the OnVUE mode, where a candidate can take the test remotely with an online proctor, or the test centre mode, where candidates take the test in a specially designed test centre. Only 64 candidates took the online version of the test (see Section 4.4).

3. Examination Results

Overall Exam Results

This section presents a comprehensive summary of the examination outcomes for the 41,354 candidates who participated in the UCAT between 7 July 2025 and 26 September 2025. From 2017 to 2021, there was consistent annual growth in the number of candidates, reflecting a period of steady expansion. A modest decline in candidate volume occurred between 2021 and 2023; however, numbers have since rebounded, reaching an all-time high in 2025.

Figure 1. Candidate Volumes since 2017



The increase observed this year is particularly notable, representing a 9% rise compared to the prior year. Relative to 2017, the candidate volume in 2025 demonstrates a significant 66% increase. This eight-year upward trajectory underscores substantial growth in participation in the UCAT. Figure 1 above visually depicts changes in candidate numbers since 2017, illustrating these trends over time.

Table 3. Cognitive Subtest and Total Scaled Score Summary Statistics

Subtest	Mean	*SD*	Min	Max
VR	602.47	80.57	300	900
DM	627.58	86.32	300	900
QR	660.83	109.98	300	900
Total	1890.87	244.84	980	2670

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 602, and the highest average score was achieved on QR with a mean of 661.

Figure 2. Subtests' Average Scaled Scores by Year since 2017

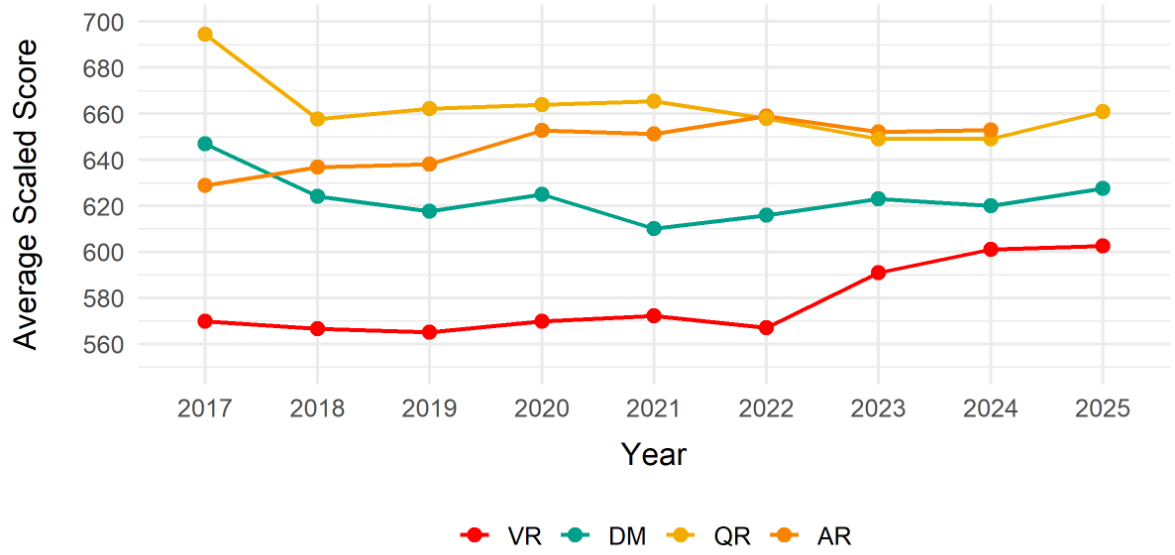


Figure 2 illustrates the progression of scaled scores from 2017 onward. The year 2017 serves as the baseline for analysis because, prior to this point, there was no operational DM section to provide comparable data. The same information is also detailed in Table 4.

Table 4. Historic Cognitive Subtests Mean Scaled Scores (2017–2025)

Subtest	2017	2018	2019	2020	2021	2022	2023	2024	2025
VR	570	567	565	570	572	567	591	601	602
DM	647	624	618	625	610	616	623	620	628
QR	695	658	662	664	665	658	649	649	661
AR	629	637	638	653	651	659	652	653	-

Between 2018 and 2021, the mean scaled scores for the VR, QR, and AR subtests remained relatively consistent. Notably, both QR and DM experienced significant decreases in their scaled scores in 2018. These drops were the result of methodological changes: QR's scaling method was revised, and DM's benchmark population was updated. During the same period, AR scores exhibited a steady upward trend, likely due to the subtest's trainability and growing familiarity among candidates and the general public.

In 2022, a timing adjustment was implemented for QR to reduce the time pressure on examinees, which was expected to raise the average scaled score. To counteract this anticipated increase, QR scores were adjusted downward by 20 points. Despite these changes, QR and AR maintained considerably higher average scaled scores than DM and VR in 2022, with VR lagging behind most noticeably.

To address the imbalance among subtest scores, further adjustments were made in 2023 and 2024. In 2023, the QR and AR scores were each reduced by 10 points, while the VR score was increased by 20 points. The same adjustments were implemented again in 2024, further reducing QR and AR scores by 10 points and increasing the VR score by another 20 points. These measures were intended to narrow the gap between subtest scores while keeping the overall total scaled score stable. The adjustments were effective in 2023 and, to a lesser degree, in 2024.

Although QR and AR were scaled down by 10 points in 2024, their average scaled scores remained nearly identical to those in 2023. This suggests that positive influences offset the downward scaling. For VR, despite a 20-point increase in scaling in 2024, the average scaled score rose by only 10 points, indicating that negative factors partially counteracted the upward adjustment.

In 2025, both VR and QR subtests were scaled down by 10 points to account for the additional minute added to the test time. Nevertheless, both subtests experienced a slight increase in their mean scaled scores. The VR mean scaled score increased by just 1 point, demonstrating that the downward scaling and extra time effectively balanced each other. In contrast, QR saw a 12-point increase in its mean scaled score, indicating a continued upward trajectory despite the downward scaling.

The DM subtest underwent the most significant changes in 2025, with the addition of 6 items and 6 minutes. Despite these modifications, the mean scaled score for DM increased by only 8 points compared to the previous year. This change is well within one standard error of measurement, implying that the subtest's difficulty remained consistent and that the additional items and time had an appropriate, balancing effect.

Looking forward, the ongoing rise in QR scores will be carefully observed, with plans to make further downward adjustments next year to maintain balanced subtest results. Overall, these changes remain well within one *SEM* for the relevant subtests, as detailed in Section 6. Statistically, such minor variations are not significant enough to warrant concern. This suggests the adjustment in scores and test structure changes are appropriately implemented such that the test scores are consistent and comparable to previous cohorts.

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

	VR	DM	QR
DM	0.66***		
QR	0.59***	0.75***	
SJT	0.49***	0.58***	0.50***

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 6. SJT Band Distribution in 2025

SJT Band	Number of Candidates	Mean Scaled Score	Percentage of Candidates	Target %
Band 1	8,852	682.98	21%	22%
Band 2	16,278	627.38	39%	38%
Band 3	12,088	557.70	29%	30%
Band 4	4,136	443.44	10%	10%
Total	41,354	600.52	100%	100%

The distribution of SJT bands in 2025, as outlined in

Table 6 above, is presented alongside the corresponding number and percentage of candidates attaining each band. Band allocations for the SJT examination are determined according to candidates' underlying scaled scores. This year, the SJT band cut-off structure was revised to promote greater stability in band distribution. The outcomes of this adjustment were notably positive, with deviations from target proportions across all bands limited to a maximum of 1%. This represents a significant improvement in consistency when compared to previous years.

Figure 3. SJT Band Proportions 2017–2025

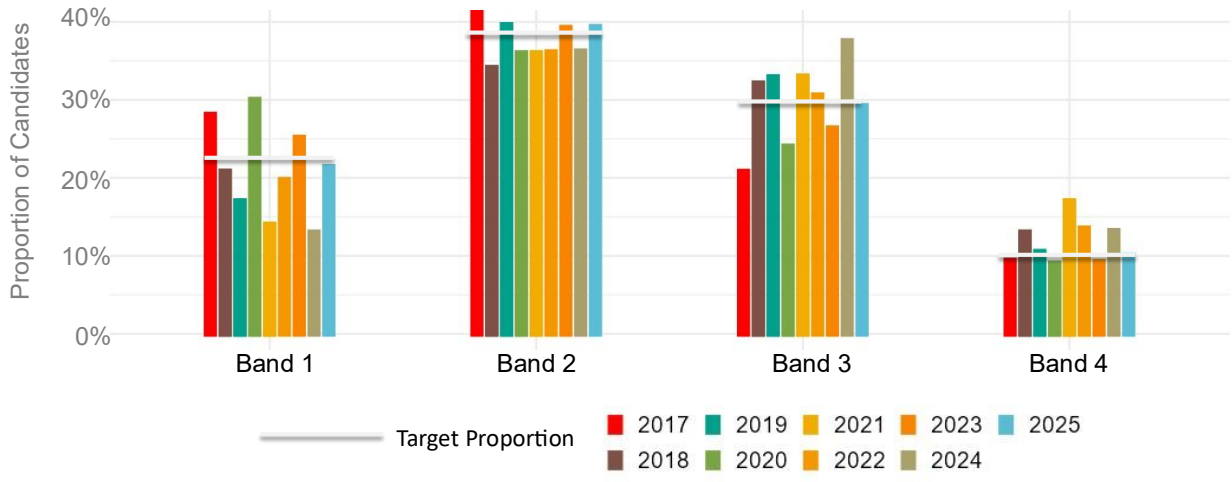


Figure 3 displays the distribution of candidates across SJT bands since 2017. Target proportions for each SJT band have been consistent for a number of years and are indicated by the white lines in Figure 3.

The distribution of SJT banding has shown significant deviations from these targets over the years, with particularly pronounced discrepancies observed in 2024. As a result, efforts were made to revise the method for determining band allocation and to minimise the cyclical patterns identified in the 2024 technical report. Following the implementation of the fixed cutoffs approach in 2025, taking into account the actual distribution of scores over a number of years, the resulting band distribution has been the closest to the targets on record, demonstrating the effectiveness of this revised methodology. It is

recommended that this approach is used for 2026 and will continue to be monitored in subsequent years to assess its ongoing effectiveness.

Special Educational Needs

There are seven exam versions available for SEN candidates who are granted extra time and breaks. However, only one candidate took the UCATSEN100SA test code. To protect their privacy, their results will not be included in most of the analyses in this technical report. Table 7 and 8 below detail the time allowances for each subtest and exam version.

Table 7. Exam Version Time Allowed

Subtest	UCAT	UCATSEN	UCATSENSA	UCATSEN50
VR	00:22:00	00:27:30	00:27:30	00:33:00
DM	00:37:00	00:46:15	00:46:15	00:55:30
QR	00:26:00	00:32:30	00:32:30	00:39:00
SJT	00:26:00	00:32:30	00:32:30	00:39:00

Table 8. Exam Version Time Allowed continued

Subtest	UCATSEN50SA	UCATSEN100	UCATSEN100SA	UCATSA
VR	00:33:00	00:44:00	00:44:00	00:22:00
DM	00:55:30	01:14:00	01:14:00	00:37:00
QR	00:39:00	00:52:00	00:52:00	00:26:00
SJT	00:39:00	00:52:00	00:52:00	00:26:00

Only 7% of candidates took a SEN version of the exam, which is similar to 2024. The most popular SEN exam was UCATSEN, as shown in Table 9 below. These exams are available to candidates who require additional time due to a special accommodation.

Table 9. Exam Version Candidate Volumes

Exam	*N*	%
UCAT	38,515	93%
UCATSEN	1,698	4%
UCATSENSA	735	2%
UCATSEN50	72	0%
UCATSEN50SA	79	0%
UCATSEN100SA	1	0%
UCATSA	254	1%
Total	41,354	100%

Historically, candidates who take a SEN version of the exam usually outperform candidates who take the non-SEN version. Table 10 summarises the scaled score statistics by exam version. SEN candidates outperformed non-SEN candidates in all three subtests. The sample sizes of UCATSEN50, UCATSEN50SA, and UCATSA are small and results for those versions should be treated with caution.

Table 10. SEN and Non-SEN Cognitive Subtests

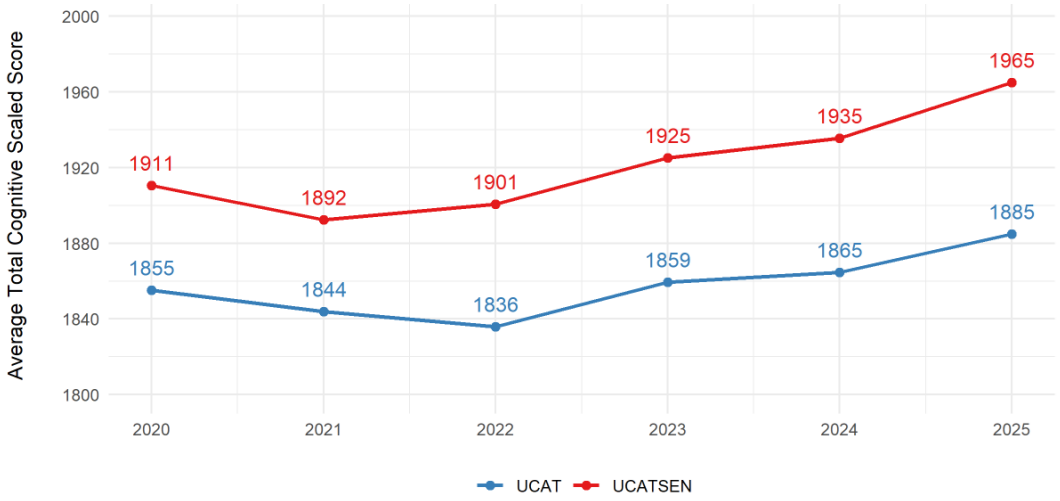
Subtest	Statistic	UCAT (38,515)	UCATSEN (1,698)	UCATSENSA (735)	UCATSEN50 (72)	UCATSEN50SA (79)	UCATSA (254)
VR	Mean	600.48	623.83	642.65	635.83	637.85	624.25
	SD	80.17	79.37	83.85	88.04	79.74	78.19
	Min	300	390	370	350	490	370
	Max	900	900	880	880	880	900
DM	Mean	625.79	650.90	656.63	641.81	645.82	648.23
	SD	86.44	79.18	85.97	79.86	88.57	75.96
	Min	300	400	390	410	430	400
	Max	900	890	890	840	870	880
QR	Mean	658.53	690.27	697.28	690.56	676.33	693.54
	SD	109.95	103.35	111.81	99.46	110.29	100.40
	Min	300	350	430	450	480	470
	Max	900	900	900	900	900	900
Total	Mean	1884.80	1964.99	1996.56	1968.19	1960	1966.02
	SD	244.78	224.61	245.60	227.25	238.04	218.34
	Min	980	1290	1200	1400	1440	1280
	Max	2670	2620	2610	2530	2540	2640

Note. UCATSEN100SA have been excluded from the table above for privacy reasons, as there were only a small number of candidates under these exam series codes.

Table 10 also shows the average total cognitive scaled score for each exam version. Overall, SEN candidates outperformed non-SEN candidates on the cognitive subtests. The mean difference in total cognitive scaled scores between UCAT and UCATSEN candidates is 80 points, which is lower than the differences observed in previous years: 109 points in 2024, 95 points in 2023, and 91 points in 2022. This reduction is due to the removal of the AR subtest from this year's exam, which resulted in lower total cognitive scaled scores and made it harder to compare cohorts across years.

Figure 4 illustrates the differences in average total cognitive scaled scores between the UCAT and UCATSEN versions after excluding the AR subtest. The cognitive subtests performance gap excluding AR between UCAT and UCATSEN candidates has progressively increased from 56 in 2020 to 80 in 2025. Continued monitoring in future years is recommended to ensure the performance gap remains stable.

Figure 4. Average Cognitive Scaled Score (AR excluded): UCAT vs UCATSEN



The pattern of SEN candidates outperforming non-SEN candidates is also evident in the SJT results. When compared to the SEN versions of the test, the non-SEN version of the exam has the lowest proportion of candidates in Band 1 and the highest proportion in Band 4. Table 11 below provides a breakdown of SJT band proportions by exam version. Results for the UCATSEN100SA versions are not disclosed for privacy reasons, as only one candidate sat this version. Among the remaining exam versions, candidates performed best on the UCATSA, where 78% of candidates were classified as either Band 1 or Band 2.

Table 11. SJT Band by Exam Version

Exam Version	Mean Scaled Score	Band 1	Band 2	Band 3	Band 4
UCAT	598.71	21%	39%	30%	10%
UCATSEN	623.30	29%	43%	24%	4%
UCATSENSA	628.35	32%	45%	20%	3%
UCATSEN50	619.92	24%	51%	18%	7%
UCATSEN50SA	627.33	30%	42%	27%	1%
UCATSA	628.35	26%	52%	19%	3%

Note. UCATSEN100SA have been excluded from the table above for privacy reasons, as there were only a small number of candidates under these exam series codes.

One of the potential concerns regarding SEN candidates is whether their higher performance is a direct result of the extra time they receive. Paton and Tiffin (2024) explored performance differences between UCAT candidates who sit standard and extended versions of the test, specifically focusing on the UCATSEN version. The study analysed data from 36,423 tests taken in 2022, including 1,612 UCATSEN tests.

The findings revealed that the higher performance of SEN candidates is not solely due to the extra time they receive. The UCATSEN group has a different sociodemographic composition compared to the UCAT group. The UCATSEN group includes more white candidates, more older candidates (20 years or older), more candidates with higher education qualifications, and fewer candidates from state schools.

When controlling for sociodemographic variables, Paton and Tiffin (2024) found that the gap in total score between UCATSEN and UCAT candidates reduces by approximately half, and the performance gap in DM and the SJT becomes non-significant. Despite remaining significant, the performance gap between UCATSEN and UCAT candidates in VR and QR dropped from 20.3 and 23.9 to 6.62 and 16.6, respectively. The performance differences substantially decreased to less than half an *SEM* after the adjustment, suggesting that a large portion of the performance differences can be explained by the inherent differences in the groups rather than the additional time provided.

It was noted in the study that after controlling for sociodemographic variables, the order of the performance gaps largely corresponds to the speededness of the subtests. QR and VR are generally considered to be relatively speeded subtests, while DM and the SJT are relatively non-speeded. This alignment between performance gaps and speededness suggests that the additional time given to UCATSEN candidates might provide a greater advantage in speeded subtests compared to less speeded ones. Consequently, efforts to minimise speededness in subtests could potentially enhance fairness between UCATSEN and UCAT candidates.

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.

Table 12. Candidates UCAS 1st Choice Distribution

UCAS 1st Choice	N	%
Medicine	22,799	55%
Dentistry	5,217	13%
Unidentifiable	13,338	32%

Table 12 presents the breakdown of candidates according to their UCAS first-choice programme. The data show that the majority of candidates, representing 55% of the total, applied for the medicine programme. This proportion of medicine applicants has remained the same as in 2024, indicating that more than half of all candidates continue to choose medicine as their preferred course. There has been a gradual decline in the proportion of medicine candidates over recent years, with figures decreasing from 59% in 2023, 63% in 2022, and 69% in 2021. This drop in candidate proportion could partially attribute to the increase of international candidates who are unidentifiable in the UCAS system. Among the candidates identifiable within the UCAS system, 81% of the candidates applied for Medicine as their first choice in 2025. The proportion of candidates applying for medicine is similar to that of 2024 and slightly lower than that observed in 2023 (82%), 2022 (86%), 2021 (88%), and 2020 (88%). Correspondingly, the

candidates applied for Dentistry program as their UCAS first-choice have gradually increased. Among the candidates identifiable in the UCAS system, 19% applied for a Dentistry programme as their first-choice, increasing from 12% from 2020. This reflects an increasing popularity of the dentistry programme among UCAT candidates.

The remaining 32% of candidates either applied for courses other than medicine or dentistry or could not be matched with UCAS data. This proportion is unchanged from 2024, but shows a slight increase from 29% in 2023, 26% in 2022, and 23% in 2021. This trend may reflect the growing number of partner international universities outside the UCAS system.

Table 13. Medicine/Dentistry Candidates: Cognitive and Total Scaled Scores

Subtest	Mean			*SD*		
	Medicine	Dentistry	None	Medicine	Dentistry	None
VR	622.21	608.60	566.31	77.60	70.42	76.85
DM	651.00	644.57	580.89	79.48	75.13	82.89
QR	688.23	686.22	604.04	105.91	100.76	97.97
Total	1961.44	1939.40	1751.24	227.47	210.49	226.20

Candidates who applied for medicine as a first choice outperformed those who applied for dentistry, as illustrated in Table 13. The highest mean scaled score was achieved on QR and the lowest on VR for both candidate groups. Candidates who did not apply for medicine or dentistry or were not matched by UCAS data performed less well than both other groups.

Better performance by medicine candidates is also evident in the SJT banding. As shown in Table 14, medicine candidates have a very slightly higher mean scaled score compared to dentistry candidates. This results in slightly more medicine candidates being classified in Band 1 than dentistry candidates, though the difference is minimal and the split is comparable.

Table 14. Medicine/Dentistry Candidates: SJT Bands

Group	Mean Scaled Score	Band 1	Band 2	Band 3	Band 4
Medicine	620.16	27%	44%	24%	4%
Dentistry	617.40	26%	45%	26%	4%
None	560.34	9%	29%	39%	23%

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.

Mode of Delivery

In 2025, the UCAT was offered in both the standard test centre and online proctored mode. Only 64 candidates took the exam in the online proctored mode, amounting to only 0.16% 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 2025, it is not possible to draw reliable inferences on differences in performance for the 2025 cohort of candidates.

Examination Results by Demographic Variables

Variation by Demographic Group

Pearson and WPG 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.

Gender

Table 15 provides the breakdown of test-takers by preferred gender term. The majority of test-takers identified as female, while only 397 indicated that they "use another term" to describe their gender or preferred not to disclose their gender.

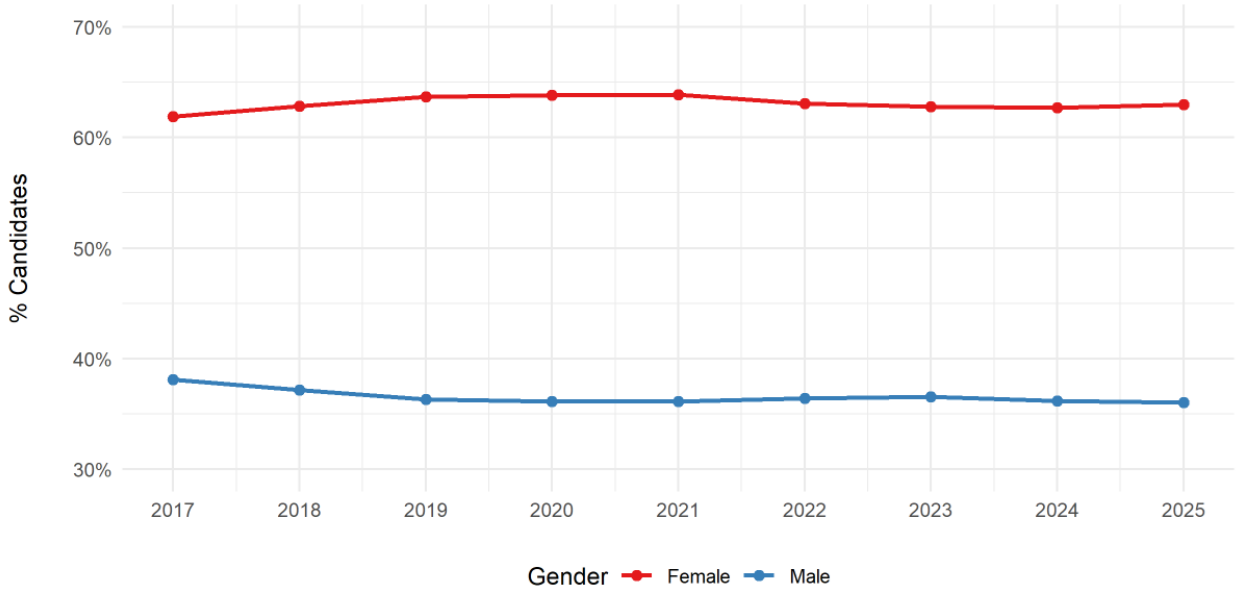
Table 15. Gender Counts

Gender	*N*	%
Female	26,049	63%

Gender	*N*	%
Male	14,908	36%
I prefer not to say	336	1%
I use another term	61	0%

The distribution of candidates by gender has remained stable since 2017, with a slight increase in female candidates from 2017 to 2019 (Figure 5).

Figure 5. Distribution of Candidates by Gender 2017–2025



Candidates who identified as male outperformed those who identified as female on all subtests except the SJT, where female candidates performed better than male candidates. Table 16 presents the differences in average scores between male and female candidates.

Table 16. Gender Scaled Scores

Subtest	Mean Scaled Score		*SD* Scaled Score	
	Female	Male	Female	Male
VR	597.01	611.15	79.75	80.63
DM	620.25	639.93	85.29	86.58
QR	646.07	686.38	105.08	113.45
Total Cognitive	1863.33	1937.45	238.65	247.71
SJT	605.62	591.40	71.84	75.22

A statistical test was used to examine whether the differences between the two groups observed in Table 16 were statistically significant. Table 17 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 that 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 male candidates and

all eligible female candidates). 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.

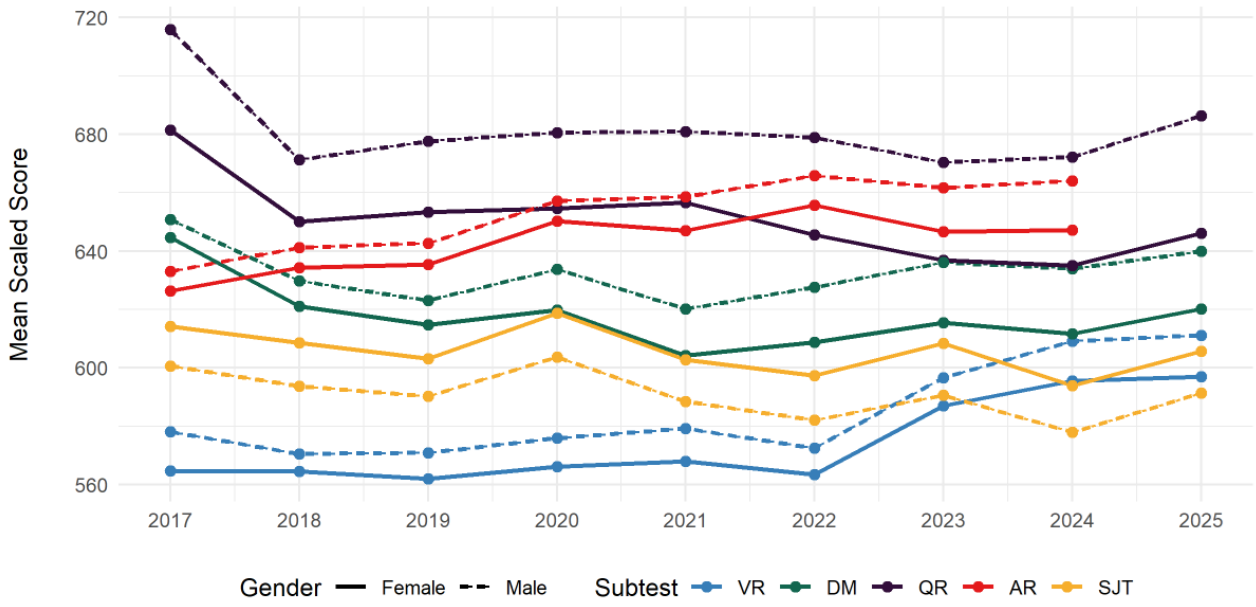
Therefore, Table 17 shows us that there are differences between male and female performance on each subtest and on the total cognitive scores, and that these differences are likely not to be the result of random chance.

Table 17. Gender t-Test

Subtest	*t*-Statistic	*df*	*p* Value
VR	17.19	40,955	< 0.01
DM	22.34	40,955	< 0.01
QR	36.27	40,955	< 0.01
Total Cognitive	29.83	40,955	< 0.01
SJT	-18.95	40,955	< 0.01

Figure 6 illustrates the subtest score differences by gender, which have remained relatively consistent year on year. Since 2017, the score gap between male and female candidates has slightly widened in the DM subtest. Additionally, since 2021, the score gap has also slightly increased in the QR subtest.

Figure 6. Scaled Score Distribution of Candidates by Gender 2017–2025



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 2022 by reducing the number of options. These options align closely with the groups used in previous reports except for UK-Chinese, which was removed as a separate category in 2022. The categories used are:

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

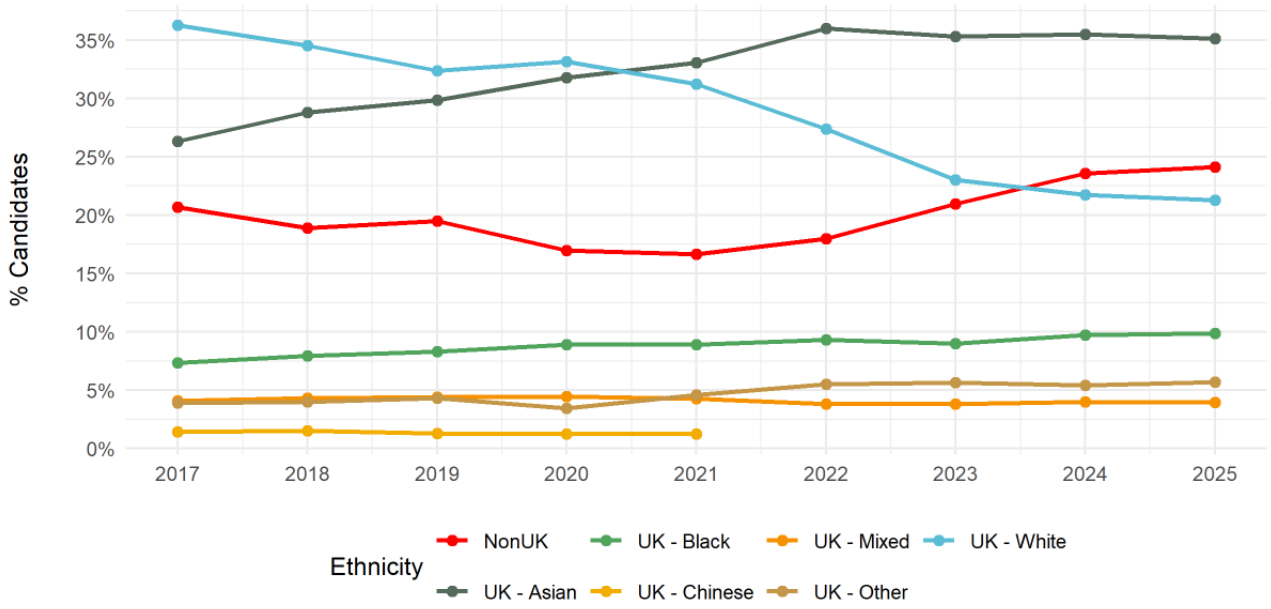
Table 18 shows the breakdown of candidates by ethnicity in the 2025 exam. The biggest candidate group was UK-Asian. Twenty-four percent of candidates were not categorised due to being non-UK candidates.

Table 18. Ethnic Group Counts

Country	Ethnic Group	*N*	% UK Candidates	% Total Candidates
UK	Asian	14,245	46%	35%
UK	White	8,627	28%	21%
UK	Black	3,992	13%	10%
UK	Other ethnic group	2,303	7%	6%
UK	Mixed	1,600	5%	4%
Non-UK	Non-UK	9,776	-	24%

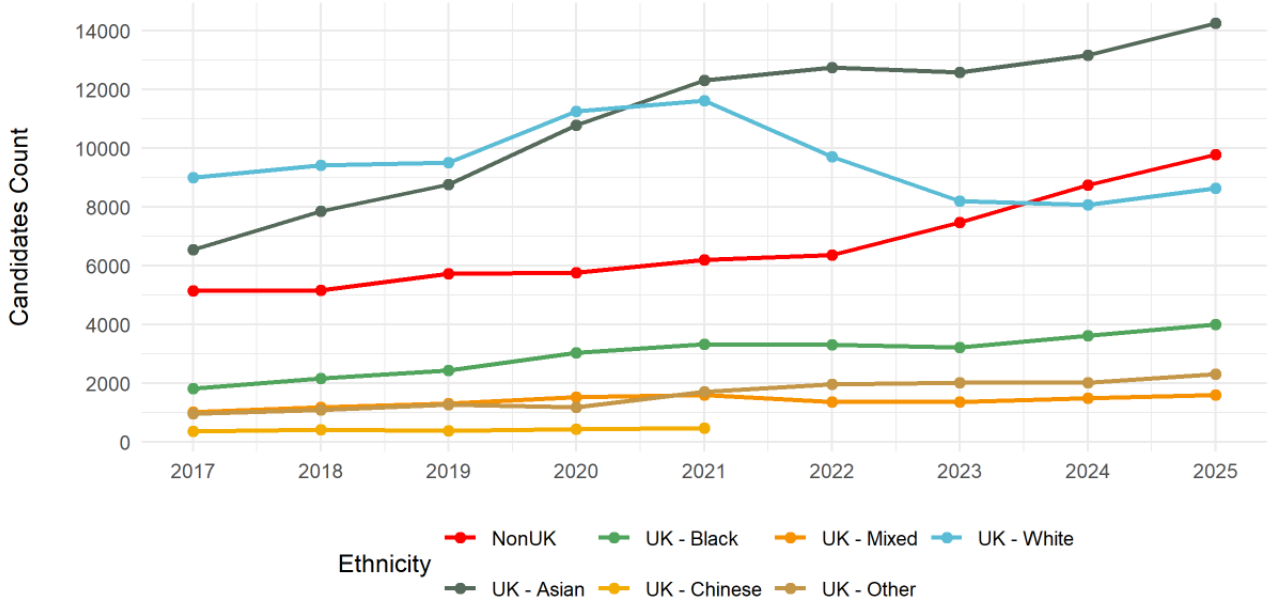
Over recent years, the proportion of candidates from most ethnic groups has remained relatively stable, with a few notable exceptions. An examination of Figure 7 reveals the evolving trends in the ethnic makeup of the candidate pool. Since 2017, there has been a gradual increase in the proportion of UK-Asian candidates, while the proportion of UK-White candidates has experienced a steady decline. A significant shift occurred in 2021, when UK-Asian candidates became the largest ethnic group in the sample, surpassing UK-White candidates. Additionally, there has been a gradual increase in the number of non-UK candidates since 2022. By 2024, the UK-White group had dropped to the third-largest ethnic group, with the non-UK group rising to the second-largest position. The distribution observed this year remains largely consistent with that of 2024, with no changes in the order of the ethnic group size represented in the candidate pool.

Figure 7. Distribution of Candidates by Ethnic Group 2017–2025



An initial review of the proportion of candidates by ethnic group may raise concerns about the decline in UK-White candidates, decreasing from over 35% in 2017 to approximately 20% in 2025. At first glance, this appears to represent a reduction of more than one third in this group. However, examining the actual number of candidates, as illustrated in Figure 8, offers greater clarity. In 2017, there were 9,003 UK-White candidates compared to 8,627 in 2025, a figure that remains relatively stable. The primary change is attributable to a significant increase in the number of UK-Asian and non-UK candidates.

Figure 8. Candidates Count by Ethnic Group 2017–2025



UK-White candidates achieved the highest average scores across all subtests compared to other ethnic groups. Table 19 provides a breakdown of the average scores for each subtest by ethnic group. UK-Black candidates had the lowest average performance in DM, QR, and the aggregated total cognitive scaled

score. For the SJT, non-UK candidates recorded the lowest average scores, while for VR, candidates from the Other Ethnic Group category achieved the lowest average scores.

Table 19. Ethnic Group Mean Scaled Score

Subtest	White	Asian	Black	Mixed	Other	Non-UK
VR	624.87	601.99	583.87	618.04	580.65	593.40
DM	653.20	629.30	597.14	645.45	612.10	616.31
QR	679.89	672.53	614.90	676.36	645.11	647.88
Total Cognitive	1957.97	1903.81	1795.91	1939.85	1837.86	1857.59
SJT	622.84	607.31	594.17	618.75	593.78	571.58

An *F*-test was used to examine whether the differences observed in Table 19 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 20 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 19 are likely to reflect true differences in performance in the candidate population.

Table 20. Ethnic Group F-Test

Subtest	*F*-Statistic	*df*	*p* Value
VR	211.92	6	< 0.01
DM	273.62	6	< 0.01
QR	230.94	6	< 0.01
Total Cognitive	285.56	6	< 0.01
SJT	460.89	6	< 0.01

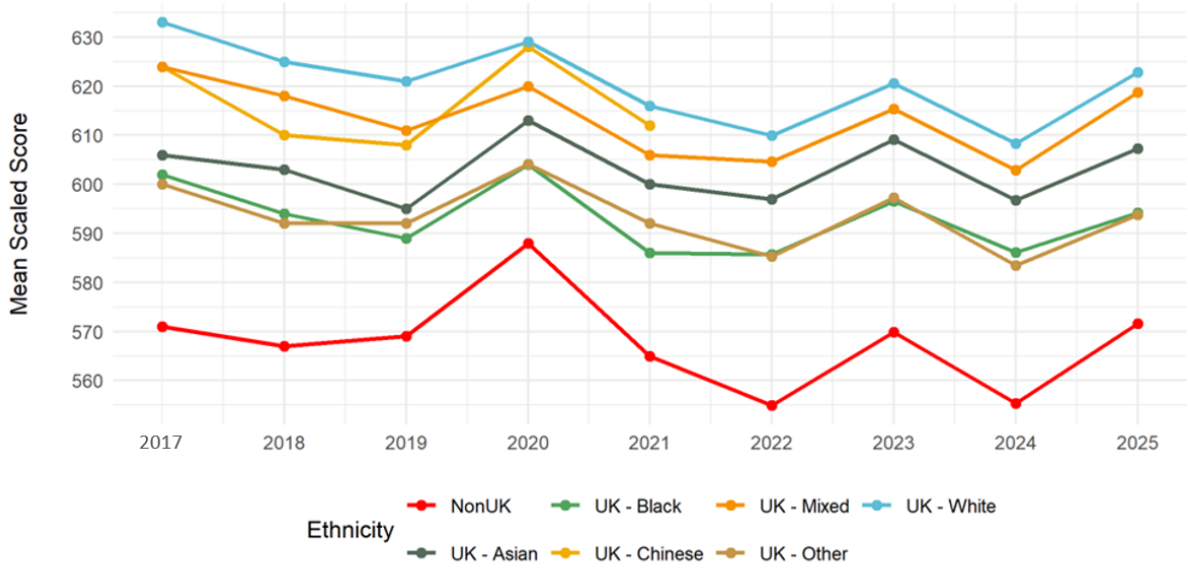
Table 21 displays the mean total cognitive scaled scores from 2020 through 2025. Owing to the discontinuation of the AR subtest, the available total cognitive scaled scores are substantially lower this year compared to previous years, rendering direct group comparisons inappropriate. Nonetheless, the ranking of ethnic groups has remained relatively consistent over time. UK-White candidates have consistently achieved the highest mean total cognitive scaled scores. They are followed by UK-Mixed, UK-Asian, non-UK, UK-Other, and UK-Black candidates. This hierarchy has demonstrated considerable stability across the assessed years, with minimal variation observed in the current reporting period.

Table 21. Mean Total Cognitive Scaled Scores from 2020

Ethnicity	2025	2024	2023	2022	2021	2020
UK - White	1,958	2,601	2,593	2,592	2,576	2,594
UK - Mixed	1,940	2,578	2,566	2,548	2,535	2,537
UK - Asian	1,904	2,543	2,520	2,503	2,491	2,488
Non-UK	1,858	2,479	2,488	2,427	2,436	2,466
UK - Other	1,838	2,464	2,455	2,433	2,433	2,428
UK - Black	1,796	2,397	2,393	2,378	2,372	2,378

Figure 9 presents the mean scaled scores for the SJT by ethnic group from 2017 to 2025. The ranking of ethnic groups on the SJT remains relatively consistent throughout this period. While the order of performance for the SJT shows minor differences compared to total cognitive scaled scores, non-UK candidates consistently achieve the lowest scores, with a notable margin separating them from other groups. This trend may be associated with situational judgement's relationship to cultural competence, as UK-based candidates are likely to possess greater familiarity with UK-specific norms and behaviours. It is important to note that all items exhibiting potential bias are systematically reviewed in the DIF section of this report and are removed if significant bias is identified. This thorough review process is conducted both during pilot testing and operational deployment, ensuring that minimal items demonstrating bias reach the operational test pool.

Figure 9. Ethnic Group Mean Scaled Score for SJT 2017–2025



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 22, 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.

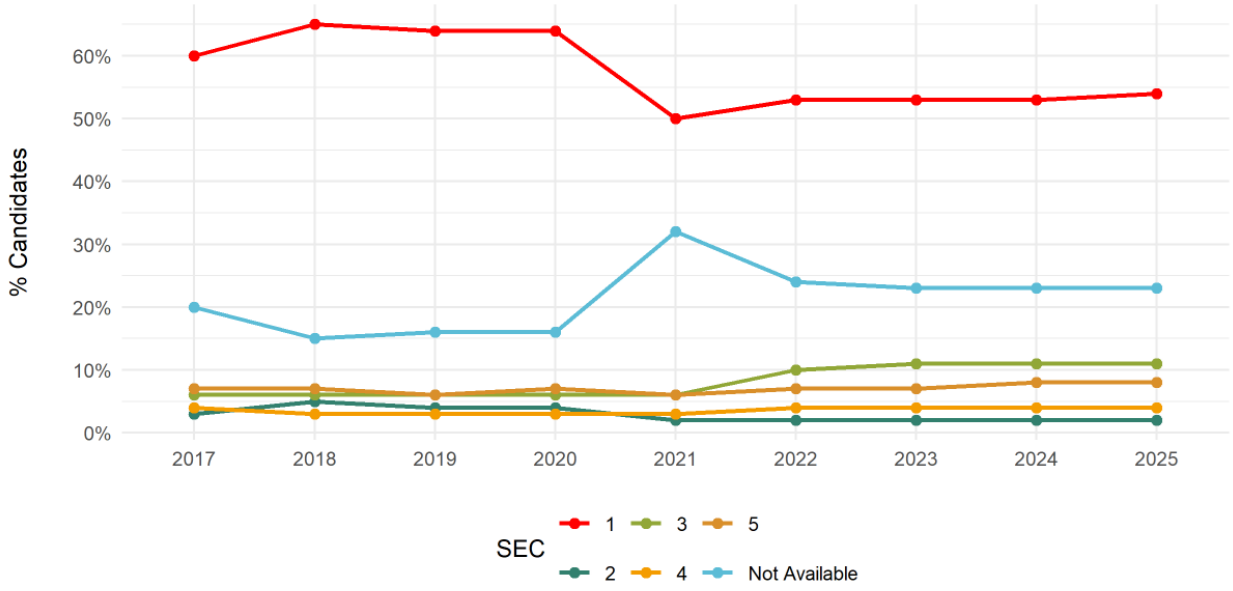
Table 22. SEC Counts

Country	SEC	*N*	% of SEC	% of All
UK	1	16,995	54%	41%
	2	568	2%	1%
	3	3,351	11%	8%
	4	1,177	4%	3%
	5	2,369	8%	6%
	Unknown	7,118	23%	17%
EU		1,357		3%
Other		8,419		20%

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
Unknown – 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 Not Available (NA) category from 2021 than in previous years, as illustrated in Figure 10. Figure 11 also 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 2025 is similar to that observed in 2024.

Figure 10. Candidates by SEC 2017–2025



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 23.

Table 23. SEC Scaled Scores

Mean Scaled Score						
Subtest	SEC 1	SEC 2	SEC 3	SEC 4	SEC 5	NA
VR	616.79	598.94	597.89	594.71	588.72	589.01
DM	645.62	617.96	623.34	619.55	609.39	610.12
QR	680.59	641.39	658.75	652.16	641.47	641.82
Total Cognitive	1,943.01	1,858.29	1,879.98	1,866.41	1,839.58	1,840.95
SJT	617.76	609.02	605.19	600.84	601.92	595.69

SD						
VR	77.35	75.46	72.42	71.65	70.39	77.83
DM	80.57	79.61	80.04	80.46	77.74	84.09
QR	105.45	98.05	101.09	100.60	97.64	104.84
Total Cognitive	229.36	219.07	219.87	220.10	212.24	235.21
SJT	62.07	67.13	65.38	66.03	66.92	72.21

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

Table 24. SEC F-Test

Subtest	*F*-Statistic	*df*	*p* Value
VR	176.39	5	< 0.01
DM	252.15	5	< 0.01
QR	183.28	5	< 0.01
Total Cognitive	263.70	5	< 0.01
SJT	130.92	5	< 0.01

Age

The majority of UCAT candidates fall within the 16–19 year age range. Only a small fraction are aged 35 or older, with an even lesser proportion under the age of 16 (see Table 25). Over recent years, there has been a consistent increase in the proportion of candidates aged 16–19; however, this trend has stabilised: 76% of candidates were in this age group in 2020, rising to 78% in 2021, 81% in 2022, and 82% in both 2023 and 2024, followed by a slight decline to 81% in 2025.

Table 25. Age Counts

Age	*N*	Percent
≤ 15	50	0%
16–19	33,533	81%
20–24	5,671	14%
25–34	1,706	4%
≥ 35	378	1%

Candidates who were aged 16–19 tended to perform better in all cognitive subtests, as illustrated in Figure 11 below. In the SJT, candidates who were 20–24 tended to perform the best. Candidates who were under 16 and over 34 had the lowest performance across all subtests 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 12.

Figure 11. Mean Scaled Scores by Age

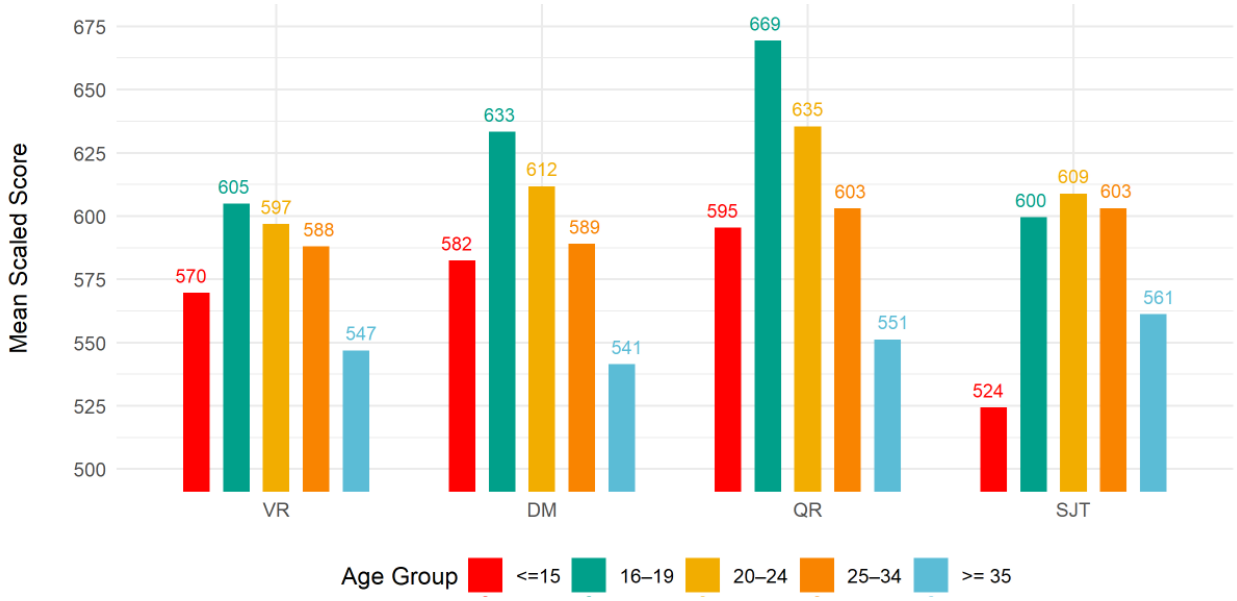
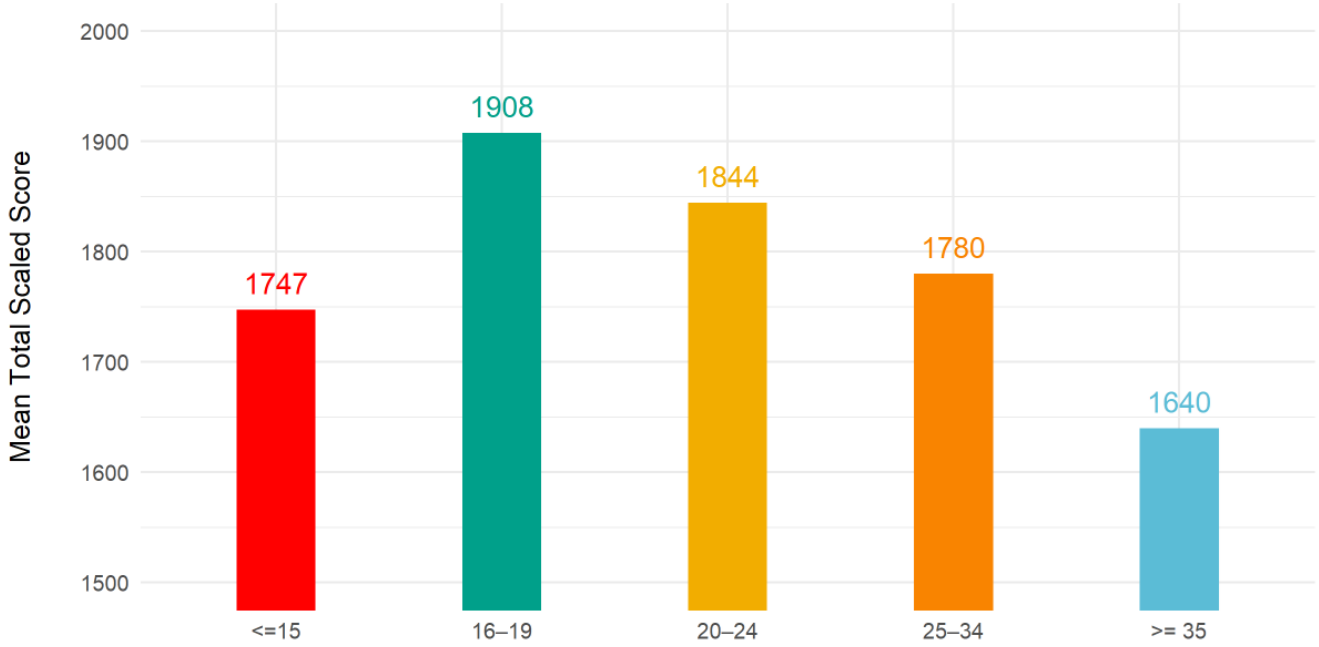


Figure 12. Mean Total Scaled Scores of Cognitive Subtests by Age



Hypothesis testing demonstrated that the differences observed among the groups is unlikely to have occurred due to chance, as shown in Table 26.

Table 26. Age F-Test

Subtest	*F*-Statistic	*df*	*p* Value
VR	75.09	4	< 0.01
DM	273.81	4	< 0.01
QR	354.05	4	< 0.01
Total	289.25	4	< 0.01
SJT	61.36	4	< 0.01

To investigate the association between age and subtest performance, Table 27 presents the correlations between candidate age and each subtest score. As indicated in the significance column, all subtests demonstrate statistically significant correlations. For the cognitive subtests, there is a slight negative correlation with age, suggesting that younger candidates generally achieved higher scores. This may reflect that most candidates undertake the test shortly after completing secondary school, whereas older candidates may have taken alternative pathways or experienced delays before entering medicine or dentistry, potentially indicating differing levels of preparedness or competitiveness at the time of testing. Nevertheless, it should be emphasised that the correlations between age and cognitive subtest performance are minimal, consistent with previous years' findings.

In contrast, whereas prior years typically showed a small positive correlation between age and SJT performance, this year reveals a very small but statistically significant negative correlation. While it had previously been hypothesised that older candidates might perform better on the SJT, the effect size observed is extremely small and close to negligible; the reversal of the correlation direction this year further suggests that this relationship is negligible and likely attributable to random variation.

Table 27. Correlation of Scaled Score with Age (ungrouped)

Subtest	Correlation	Significance
VR	-0.09	$p < 0.01$
DM	-0.16	$p < 0.01$
QR	-0.18	$p < 0.01$
Total Cognitive	-0.17	$p < 0.01$
SJT	-0.01	$p = 0.014$

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

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 2025 had a school leaver qualification (83%), 15% had a degree or above, and a small minority had no formal qualifications. These are consistent with what was observed in 2024.

Candidates with school leaver qualifications performed better on average on all cognitive subtests and the total cognitive scaled score. Candidates with a degree or above performed better on average on the SJT, as shown in Table 28. Table 29 shows that the differences observed in Table 28 are statistically significant.

Table 28. Education Scaled Scores

Subtest	School Leaver Qualification	Degree Level or Above
N	34,205	6,265
Mean Scaled Score		
VR	604.77	594.37
DM	632.71	604.91
QR	668.65	624.38
Total Cognitive	1,906.13	1,823.66
SJT	600.78	605.93
SD		
VR	80.09	81.10
DM	85.38	85.78
QR	109.96	100.68
Total Cognitive	243.31	235.71
SJT	71.84	75.76

Table 29. Education t-Test

Subtest	*t*-Statistic	*df*	*p* Value
VR	-9.43	40,468	< 0.01
DM	-23.68	40,468	< 0.01
QR	-29.67	40,468	< 0.01
Total Cognitive	-24.78	40,468	< 0.01
SJT	5.17	40,468	< 0.01

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 30 below.

Table 30. Candidate Count by Residence

Country of Permanent Residence	*N*	Percent
UK	31,578	76%
Rest of World	8,419	20%
EU	1,357	3%

As in previous technical reports, candidates from the EU and the Rest of the World are combined into a single category referred to as Non-UK. Since 2022, the proportion of candidates residing outside the UK has shown a slight, gradual increase, possibly due to the growing number of international partner universities, as illustrated in Figure 13.

Figure 13. Country of Residence 2017–2025

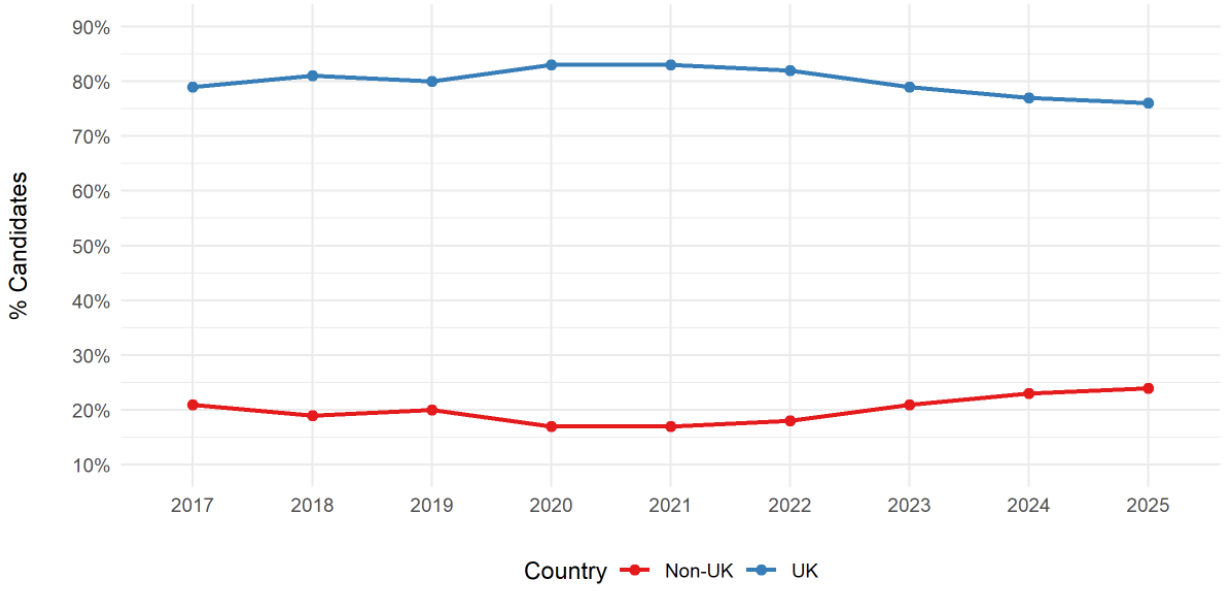


Table 31 demonstrates that UK candidates outperform those from EU and Rest of the World across all subtests. Candidates from the Rest of the World achieve higher scores than EU candidates in DM and QR, while EU candidates perform better in VR and SJT.

Table 31. Candidate Scaled Scores by Residence

Subtest	UK	Rest of World	EU
Mean Scaled Score			
VR	605.27	593.35	593.66
DM	631.07	616.60	614.53
QR	664.84	652.59	618.69
Total Cognitive	1,901.18	1,862.54	1,826.88
SJT	609.47	569.16	586.60
SD			
VR	77.26	91.54	78.45
DM	82.69	98.01	84.81
QR	105.47	125.54	96.88
Total Cognitive	232.64	284.99	228.20
SJT	66.11	88.83	73.78

An *F*-test of the differences observed between UK and non-UK candidates is presented in Table 32 below. It shows that the differences are statistically significant.

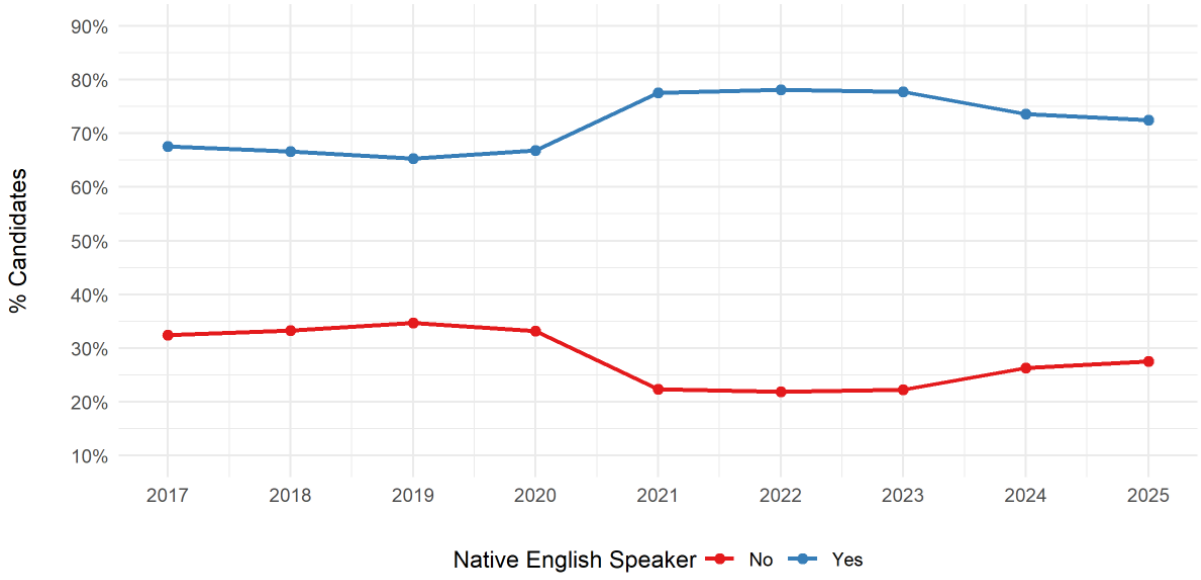
Table 32. Residence F-Test

Subtest	*F*-Statistic	*df*	*p* Value
VR	81.47	2	< 0.01
DM	109.97	2	< 0.01
QR	145.22	2	< 0.01
Total Cognitive	131.52	2	< 0.01
SJT	1,080.86	2	< 0.01

First Language

In 2025, the majority of candidates who sat the UCAT reported that English was their first or primary language. Since 2017, the proportion of candidates indicating English as their first or primary language has fluctuated (Figure 14). Between 2023 and 2025, there was a slight decrease in the proportion of candidates with English as their first or primary language. It is worth noting that the change observed in 2021 is due to a minor adjustment to the wording of this question.

Figure 14. Count of Language 2017–2025



Across all subtests, candidates who stated that English was their first or primary language outperformed those who stated that English was not their first or primary language regardless of their country of residence, as shown in Table 33 below.

Table 33. Scaled Scores by Language and Country of Residence

Subtest	Country of Residence	First Language	*N*	% of *N*	Mean	*SD*
VR	UK	English	24,653	60%	613.13	75.62
		Other	6,925	17%	577.31	76.57
	non-UK	English	5,323	13%	622.97	85.16
		Other	4,453	11%	558.04	82.17

Subtest	Country of Residence	First Language	*N*	% of *N*	Mean	*SD*
DM	UK	English	24,653	60%	639.05	80.13
		Other	6,925	17%	602.64	85.36
	non-UK	English	5,323	13%	642.02	89.76
		Other	4,453	11%	585.58	94.82
QR	UK	English	24,653	60%	672.53	104.03
		Other	6,925	17%	637.45	106.01
	non-UK	English	5,323	13%	674.26	119.45
		Other	4,453	11%	616.35	118.66
Total Cognitive	UK	English	24,653	60%	1,924.71	225.94
		Other	6,925	17%	1,817.39	236.83
	non-UK	English	5,323	13%	1,939.25	262.42
		Other	4,453	11%	1,759.97	264.47
SJT	UK	English	24,653	60%	614.63	62.05
		Other	6,925	17%	591.12	76.09
	non-UK	English	5,323	13%	594.82	71.50
		Other	4,453	11%	543.81	95.54

In line with the other demographic categories, a test was carried out to understand whether the differences observed in Table 33 can be considered statistically significant. Table 34 shows that such differences between the two groups are unlikely to have occurred by chance.

Table 34. Language t-Test

Subtest	*t*-Statistic	*df*	*p* Value
VR	52.51	41,352	< 0.01
DM	47.10	41,352	< 0.01
QR	36.62	41,352	< 0.01
Total Cognitive	50.60	41,352	< 0.01
SJT	49.00	41,352	< 0.01

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 status and qualification on QR and VR, meaning SEN had an effect 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 likely to be 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 condensed analysis of variance was also performed this year to continue tracking performance differences between UCAT and UCATSEN candidates. The findings are presented in Table 35. After controlling for demographic variables (refer to the note in Table 35), exam version remained a significant factor, with UCATSEN candidates outperforming their UCAT counterparts. The most pronounced difference was noted in the QR subtest. Notably, in both 2023 and 2024, the QR subtest exhibited the smallest difference among all subtests, whereas the 2022 analysis identified it as having the largest disparity. There has been an overall reduction in speededness in the QR subtest over recent years; however, the relationship between this decrease and fluctuations in effect size remains unclear. Ongoing monitoring will be implemented to determine if definitive conclusions can be drawn as additional data becomes available.

Table 35. Subtest Performance Differences: UCAT and UCATSEN (controlling for demographic variables)

Subtest	*F*	*p*	η^2
VR	122.42	<.0001	0.0028
DM	155.89	<.0001	0.0036
QR	185.39	<.0001	0.0042
SJT	85.34	<.0001	0.0020

Note. The comparison was only made between UCAT and UCATSEN exam codes, which accounted for 99% of the candidates. The other 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.

4. Exam Timing Analysis

The section time for each candidate is calculated by summing the item and review time for each item and candidate. Table 36 shows the exam timing for each version of the UCAT.

Table 36. Subtest Section Timing: Non-SEN and SEN

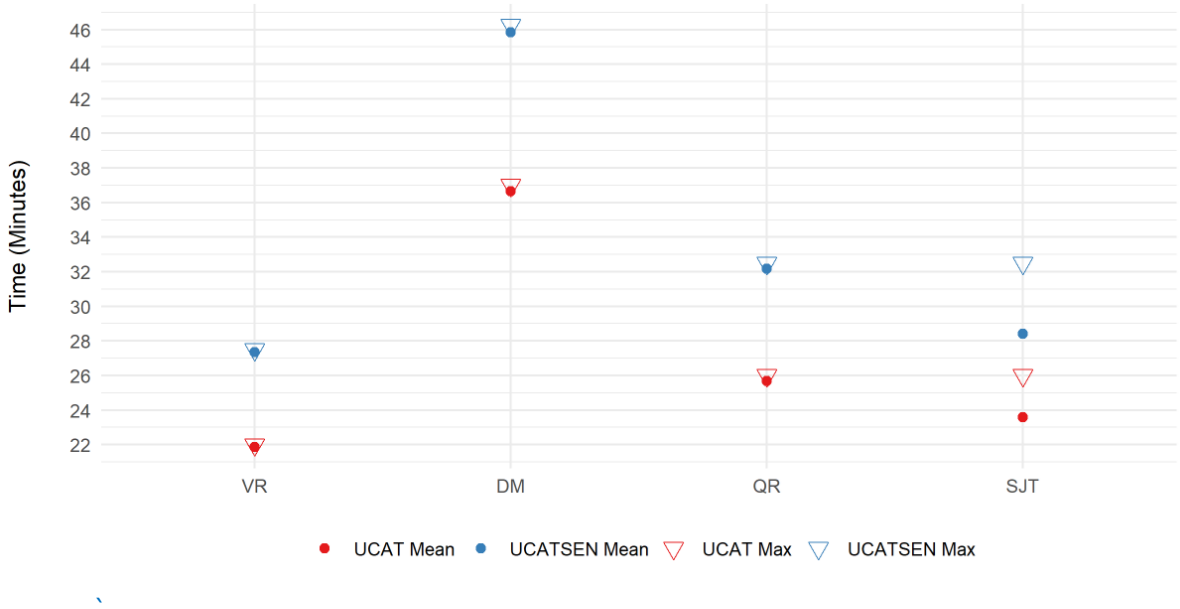
Statistic	Subtest	UCAT (38,515)	UCATSEN (1,698)	UCATSENSA (735)	UCATSEN50 (72)	UCATSEN50SA (79)	UCATSA (254)
Mean	VR	00:21:51	00:27:21	00:27:21	00:32:51	00:32:47	00:21:49
	DM	00:36:39	00:45:51	00:45:49	00:54:31	00:55:11	00:36:35
	QR	00:25:40	00:32:11	00:32:06	00:38:34	00:38:41	00:25:41
	SJT	00:23:35	00:28:24	00:27:23	00:31:42	00:32:59	00:23:05
SD	VR	00:00:30	00:00:24	00:00:20	00:00:14	00:00:33	00:00:32
	DM	00:01:11	00:01:06	00:01:33	00:03:25	00:00:28	00:01:20
	QR	00:01:18	00:00:55	00:01:23	00:01:15	00:00:42	00:01:15
	SJT	00:03:25	00:05:05	00:05:54	00:07:30	00:07:13	00:03:55
Min	VR	00:00:32	00:21:38	00:24:19	00:31:48	00:29:05	00:17:30
	DM	00:00:37	00:30:31	00:15:42	00:28:16	00:53:08	00:22:52
	QR	00:00:28	00:15:45	00:12:56	00:31:55	00:34:38	00:08:16
	SJT	00:01:02	00:09:34	00:02:40	00:15:21	00:18:08	00:09:17
Max	VR	00:22:00	00:27:30	00:27:30	00:33:00	00:33:00	00:22:00
	DM	00:37:00	00:46:15	00:46:15	00:55:30	00:55:30	00:37:00
	QR	00:26:00	00:32:30	00:32:30	00:39:00	00:39:00	00:26:00
	SJT	00:26:00	00:32:30	00:32:30	00:39:00	00:39:00	00:26:00

Note. UCATSEN100SA has been excluded from the table above for privacy reasons, as there were only a small number of candidates under this exam series code.

There is no general consensus on how to define speededness operationally. One approach is to assess it by examining how closely the average time candidates spend on a subtest approaches the total time allowed, as shown in Table 36. The cognitive subtests of the UCAT version are considered quite speeded, as the mean time spent on each subtest is close to the maximum time allowed, except for the SJT, which is notably less speeded.

The SEN versions of the exam are slightly less speeded than the UCAT version. However, the difference between the UCAT and UCATSEN versions—the latter being the only SEN version with enough candidates for reliable comparison—is minimal, as illustrated in Figure 15. For both UCAT and UCATSEN, the difference between the average time used and the maximum time allowed is almost negligible for the cognitive subtests, and the difference is more noticeable for the SJT subtest.

Figure 15. Mean and Maximum Time for UCAT and UCATSEN



Test timing is analysed in greater detail in Table 37. The data indicate that the most time-pressured subtests for the non-SEN exam version are VR and QR, where 87% and 90% of candidates, respectively, completed all items, while 6% and 4% did not attempt five or more items. These figures represent an improvement compared to the previous year, which could be attributed to the additional minute allocated to both VR and QR subtests in 2025. Among all subtests, the SJT remains the least affected by time constraints.

Table 37. Subtest Section Timing: Non-SEN and SEN UCAT Incomplete Tests

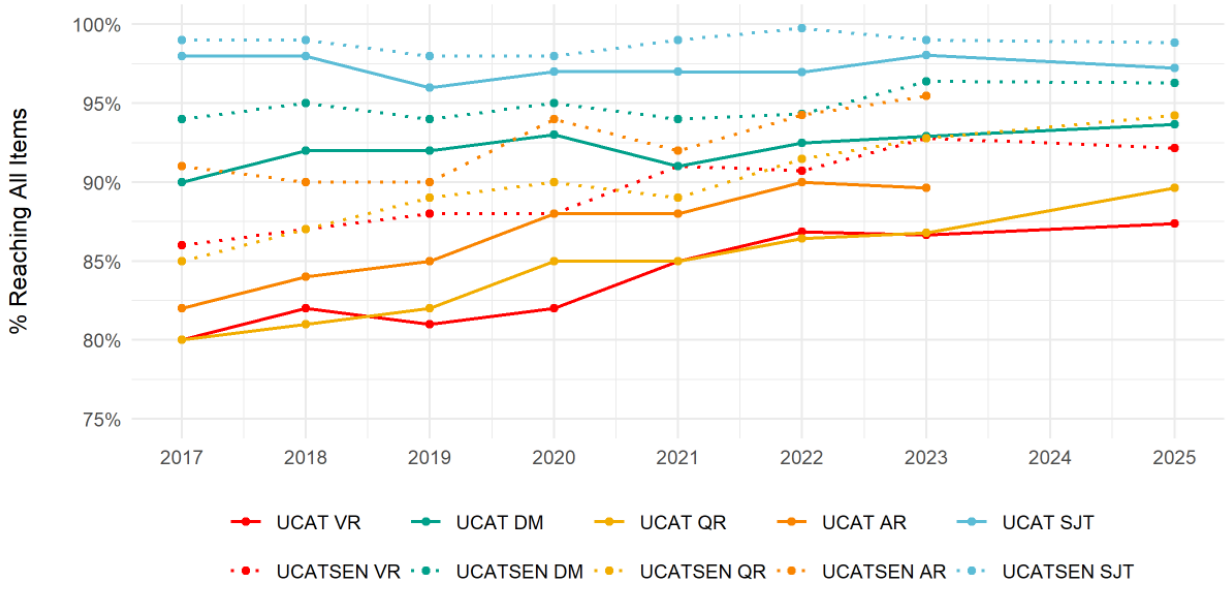
Exam	Subtest	Reached All Items *N*	Reached All Items %	Five or More Items Unreached *N*	Five or More Items Unreached %	Mean Number of Unreached Items for Incomplete Tests Only
UCAT	VR	33,658	87%	2,385	6%	6.58 (4857)
	DM	36,073	94%	824	2%	4.34 (2442)
	QR	34,520	90%	1,655	4%	5.3 (3995)
	SJT	37,452	97%	218	1%	3.95 (1063)
UCATSEN	VR	1,565	92%	53	3%	5.62 (133)
	DM	1,635	96%	18	1%	3.71 (63)
	QR	1,600	94%	34	2%	4.4 (98)
	SJT	1,678	99%	4	0%	4.2 (20)
UCATSENSA	VR	666	91%	22	3%	5.04 (69)
	DM	701	95%	12	2%	3.97 (34)
	QR	699	95%	13	2%	4.25 (36)
	SJT	726	99%	2	0%	5.44 (9)
UCATSEN50	VR	69	96%	1	1%	8 (3)
	DM	70	97%	1	1%	6.5 (2)
	QR	69	96%	0	0%	1.33 (3)
	SJT	72	100%	0	0%	0 (0)
UCATSEN50SA	VR	74	94%	1	1%	4.4 (5)
	DM	77	97%	0	0%	1.5 (2)

Exam	Subtest	Reached All Items *N*	Reached All Items %	Five or More Items Unreached *N*	Five or More Items Unreached %	Mean Number of Unreached Items for Incomplete Tests Only
	QR	74	94%	2	3%	4.4 (5)
	SJT	79	100%	0	0%	0 (0)
UCATSA	VR	240	94%	6	2%	6.43 (14)
	DM	243	96%	2	1%	3.73 (11)
	QR	243	96%	2	1%	4.55 (11)
	SJT	250	98%	1	0%	3 (4)

Note. UCATSEN100SA have been excluded from the table above for privacy reasons, as there were only a small number of candidates under these exam series codes.

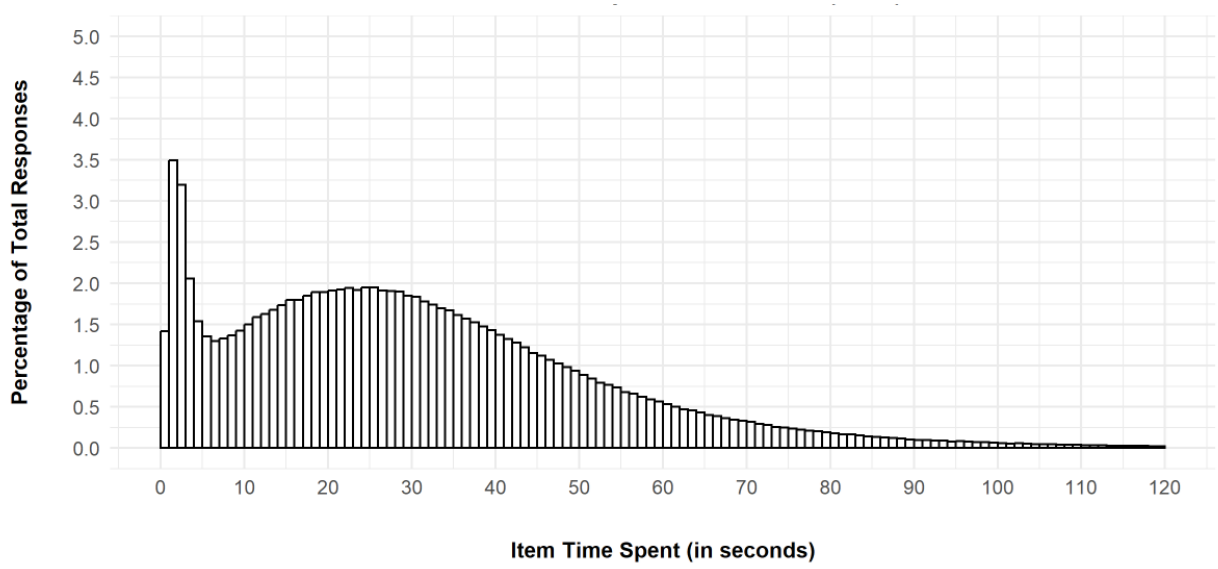
The test is being actively updated to reduce its speededness. Figure 16 illustrates the percentage of candidates reaching all items since 2017. Over this period, VR, QR, and AR have become less speeded, while DM and SJT have fluctuated within a fairly narrow band and have remained relatively non-speeded.

Figure 16. Candidates Reaching All Items 2017–2025



In 2025, after AR was removed, VR and QR each received an extra minute to reduce speededness, and DM was restructured with more items and time. QR's completion rate improved significantly, while VR's increase was minor and similar to 2022. The DM subtest maintained a comparable completion rate, indicating the restructure did not negatively impact speededness.

Figure 17. VR Response Time Distribution – 2025



The factor of guessing has been considered when evaluating speededness since 2022.

Figure 17 to Figure 23 illustrate the distribution of item response times for the four subtests. With a large sample size, these distributions are theoretically expected to follow a unimodal curve. However, bimodal distributions in the VR, DM, and QR subtests suggest the presence of two distinct behavioural patterns. The left-hand peak (local maximum), centred around 2–3 seconds with a narrow spread, contrasts with the broader peak (local maximum) on the right-hand side. The left-hand peak likely reflects rushed guessing behaviour, as it is highly unlikely that any item type could be completed in such a short time. The right-hand peak, by contrast, likely represents the actual time spent on non-guessed items. The valley (local minimum) between these peaks represents the overlap of the two distributions.

Examining the VR item response time distribution for 2025 reveals a distinctive bimodal pattern. The initial peak in the distribution occurs for responses given between 1 and 2 seconds, accounting for 3.5% of all responses. This is closely followed by responses given between 2 and 3 seconds, which comprise approximately 3.25% of total responses.

When comparing these findings to historical data, a similar pattern emerges in 2021, where the left-hand peak at 1 to 2 seconds also represented around 3.5% of responses. However, between 2022 and 2024, there was a gradual increase in the proportion of rapid responses. In 2024, responses within 1 to 2 seconds made up about 4.75% of the total, as illustrated in Figure 18. This trend indicates that, during this period, more candidates were compelled to respond within a very short timeframe, suggesting an increase in guessed responses and highlighting the heightened speededness of the test.

The situation began to improve in 2025, possibly attributed to the introduction of an additional minute for the VR subtest. This adjustment appears to have alleviated the pressure on candidates, resulting in the response time distribution reverting to levels observed in 2021. The decline in rapid, likely guessed,

responses suggests a reduction in test speededness and a more manageable pacing for candidates.

Figure 18. VR Response Time Distribution – 2021 to 2025

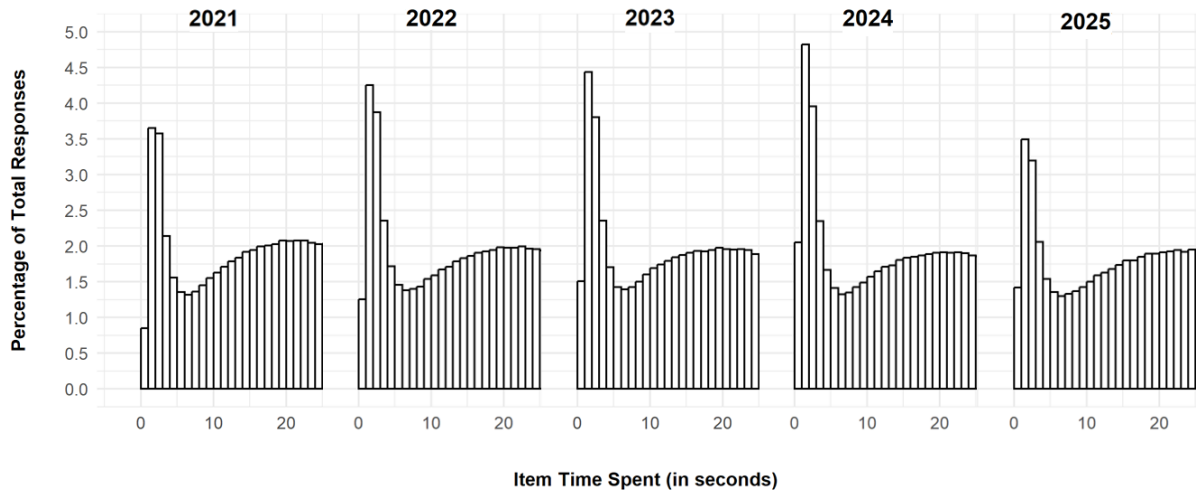


Figure 19 illustrates the response time distribution for the DM subtest in 2025. Unlike VR and QR, the height of the first peak, representing responses made between 2 and 3 seconds, is relatively small, accounting for approximately 0.9% of total responses. The majority of responses fall within the main underlying distribution of the bimodal curve, with over 97% of responses made in more than 5 seconds. This reflects DM as a relatively less speeded subtest, resulting in a lower proportion of guessed responses. Figure 20 shows the response time distributions for DM from 2021 to 2025. While there are minor fluctuations and changes across the years, the distributions remain largely consistent over time.

Figure 19. DM Response Time Distribution – 2025

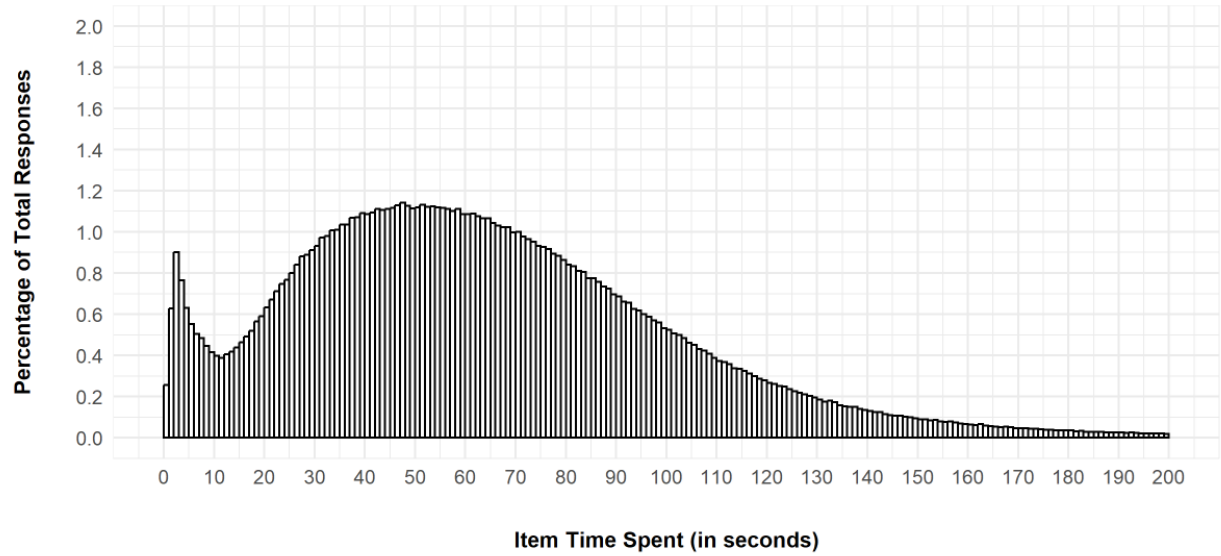


Figure 20. DM Response Time Distribution – 2021 to 2025

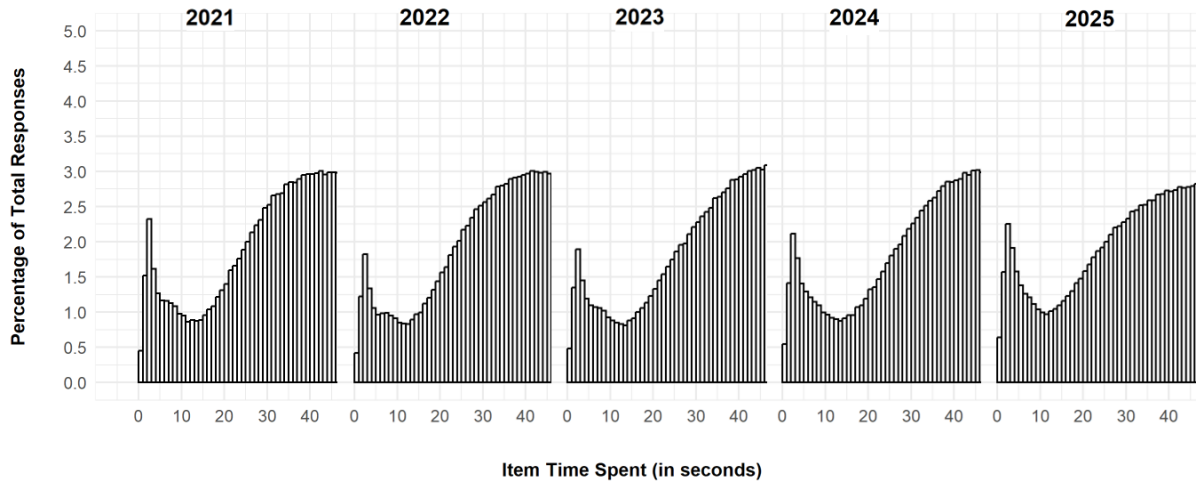
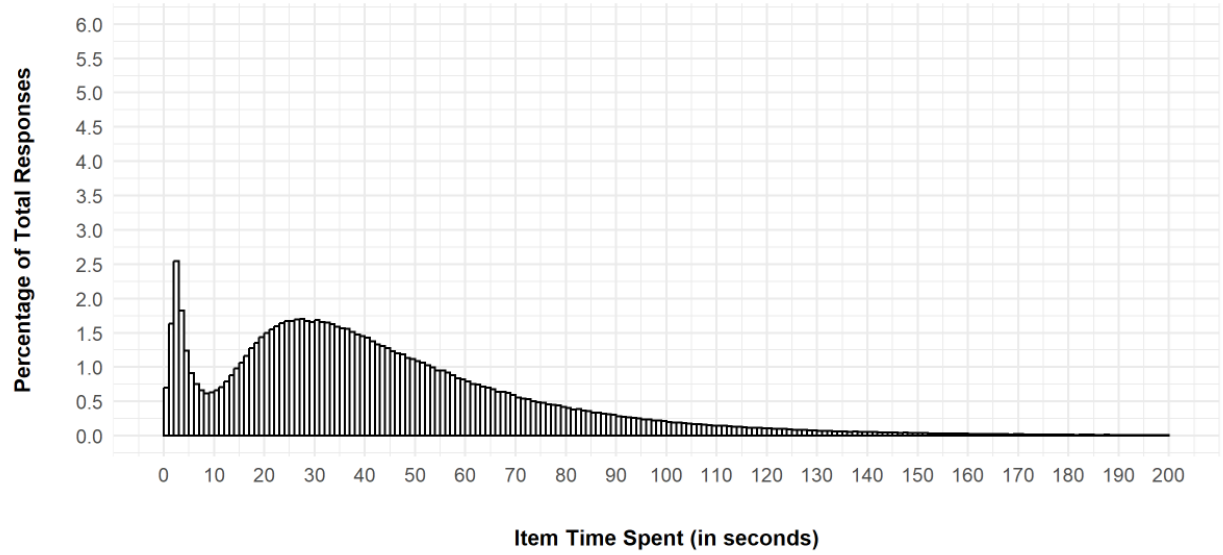
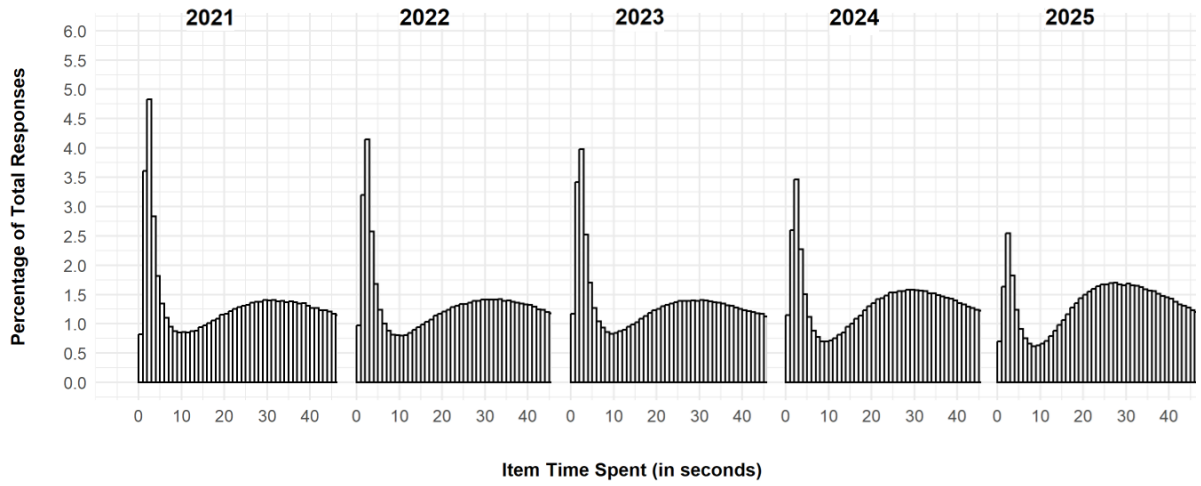


Figure 21. QR Response Time Distribution – 2025



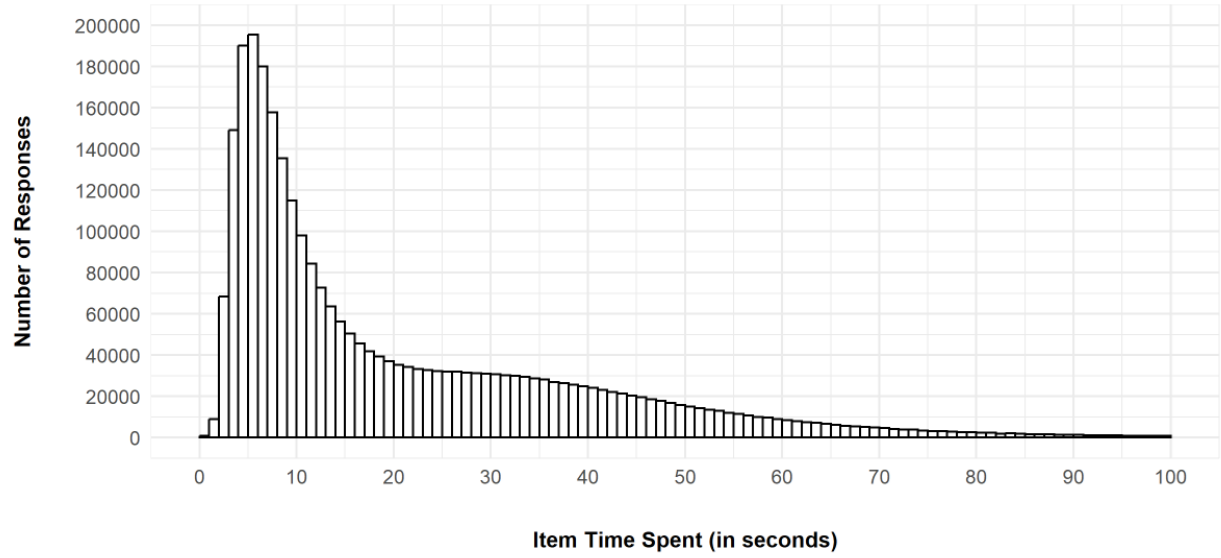
The response time distribution for QR is relatively similar to VR, with the first peak of the bimodal distribution significantly higher than the second peak. Figure 21 shows the QR response time distribution in 2025. The first peak, representing guessed responses between 2 to 3 seconds, accounts for approximately 2.5% of total responses. Figure 22 illustrates the QR distributions from 2021 to 2025, showing an opposite trend to VR. The first peak (guessed responses) decreased from around 4.8% in 2021 to 2.5% in 2025. This indicates an improvement in the speededness of the QR subtest over time.

Figure 22. QR Response Time Distribution – 2021 to 2025



In contrast to VR, DM, and QR, the SJT subtest displays skewed unimodal distributions, as shown in Figure 23. This is likely due to low item response times overlapping with guessed responses. This pattern makes it difficult to assess speededness based on a distinct clustered peak of guessed responses, as separating guessed from non-guessed responses could be complicated. Consequently, a similar examination was not conducted for these subtests.

Figure 23. SJT Response Time Distribution – 2025



By excluding responses shorter than the valley duration, it is possible to filter out most of the guessed responses, along with some rapidly answered non-guessed responses. This method provides a practical way to estimate speededness for the VR, DM, and QR subtests by discounting guessed responses. Further examination of speededness for the VR, DM, and QR subtests involved excluding responses based on various guessing thresholds. The choice of threshold is relatively subjective and produces different outcomes. A 1-second threshold, used in previous years, primarily excluded only the most hasty responses. A 5-second threshold effectively removed the peak and responses below the peak of the guessing distribution, eliminating most guessed responses while also excluding a small portion of overlapping non-guessed responses. A 10-second

threshold, which surpasses the valley for both VR and QR and approximates that of DM, likely filtered out nearly all guessed responses but also excluded a notable number of non-guessed responses.

The overlapping distributions of guessed and non-guessed responses in SJT make applying a fixed threshold less effective and may inadvertently exclude a significant number of non-guessed responses. Hence, the similar analysis for SJT is intentionally omitted in Table 38 to avoid unnecessary confusion.

Table 38. 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	88%	98%	100%	100%	100%
	Excluding responses ≤ 1s	71%	97%	98%	100%	100%
	Excluding responses ≤ 5s	19%	87%	80%	91%	98%
	Excluding responses ≤ 10s	2%	80%	73%	82%	91%
DM	All responses included	94%	99%	100%	100%	100%
	Excluding responses ≤ 1s	90%	99%	100%	100%	100%
	Excluding responses ≤ 5s	61%	96%	94%	100%	100%
	Excluding responses ≤ 10s	43%	94%	91%	97%	100%
QR	All responses included	90%	99%	100%	100%	100%
	Excluding responses ≤ 1s	82%	98%	100%	100%	100%
	Excluding responses ≤ 5s	41%	91%	86%	97%	100%
	Excluding responses ≤ 10s	26%	87%	81%	92%	100%
SJT results are intentionally omitted to avoid confusion						

Employing a balanced 5-second exclusion threshold, the proportion of candidates completing all items in VR, DM, and QR without guessing decreased significantly to 19%, 61%, and 41%, respectively. These results indicate that a relatively small subset of candidates were able to complete these subtests within the allotted time without resorting to guesses. Nevertheless, on average, candidates attempted 87% of VR items, 96% of DM items, and 91% of QR items without guessing, demonstrating a high level of engagement across most test components. Irrespective of the guessing exclusion, VR and QR continued to be the most speeded subtests, with VR showing a marginally higher degree of speededness compared to QR. In line with overall completion rates, VR exhibited a modest increase in 2025, while QR showed a marked improvement for the same period. Specifically, 41% of candidates completed QR without guessing in 2025, up from 27% in 2024 and 20% in 2023, reflecting notable progress in QR speededness. In contrast, VR experienced only a slight enhancement despite an additional minute allotted, with completion rates at 19% in 2025, 13% in 2024, and 14% in 2023.

5. Test Form Analysis

Table 39 shows the number of candidates who received each form. Candidates who were eligible for extra time and/or special accommodations were assigned either Form 1 or Form 2.

Table 39. Candidates by Form

Form	Candidates
Form 1	8,668
Form 2	8,568
Form 3	8,730
Form 4	7,753
Form 5	7,635

Table 40 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.

Table 40. Cognitive Raw Score Test Statistics

Subtest	Form	Mean	*SD*	Min	Max	Alpha	*SEM*
VR (40 items)	Form 1	22.19	6.09	2	39	0.78	2.86
	Form 2	22.33	6.06	3	40	0.77	2.91
	Form 3	22.57	6.36	3	40	0.79	2.91
	Form 4	22.04	6.17	2	40	0.78	2.89
	Form 5	22.1	6.18	3	40	0.77	2.96
DM (31 items)	Form 1	23.09	6.66	1	40	0.78	3.12
	Form 2	21.77	6.7	1	41	0.78	3.14
	Form 3	22.79	6.92	1	42	0.81	3.02
	Form 4	22.25	7.1	3	42	0.81	3.09
	Form 5	22.85	6.58	2	40	0.77	3.16
QR (32 items)	Form 1	21.46	6.98	0	32	0.89	2.32
	Form 2	21.35	6.88	0	32	0.88	2.38
	Form 3	22	6.88	0	32	0.9	2.18
	Form 4	20.9	6.8	1	32	0.88	2.36
	Form 5	20.41	6.69	0	32	0.87	2.41

Table 40 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

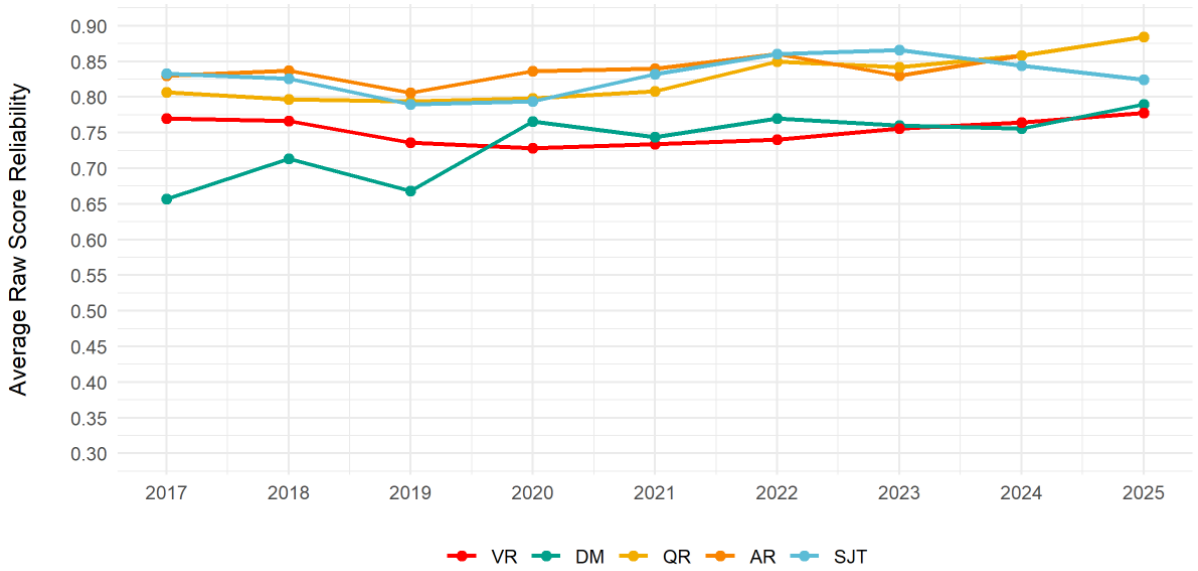
additional column called "mean percent raw score" is added (Table 41). Similar to the cognitive results, the reliability is adequately high and the *SEM* adequately low for the SJT.

Table 41. SJT Raw Score Test Statistics (246 score points)

Form	Mean	*SD*	Min	Max	Mean Percent Raw Score	Alpha	*SEM*
Form 1	183.93	19.81	69	224	74.77%	0.82	8.40
Form 2	182.69	20.48	49	229	74.26%	0.83	8.44
Form 3	181.25	20.49	49	232	73.68%	0.82	8.69
Form 4	182.74	20.23	56	227	74.28%	0.82	8.58
Form 5	182.22	21.49	43	226	74.07%	0.83	8.86

Figure 24 presents the average Cronbach's alpha for each subtest in every form since 2017. Prior to 2019, this metric reflects the mean of three forms, while from 2019 onwards, it represents the mean of five forms. The reliability of DM has increased since its introduction in 2017. In contrast, VR experienced a slight decline in reliability between 2019 and 2022 but subsequently returned to levels observed in 2017, maintaining overall consistency over time. QR has shown continuous improvement in reliability throughout the years, whereas SJT exhibited an upward trend, with a minor decrease in the past two years.

Figure 24. Raw Score Reliability 2017–2025



Raw scores are scaled and reported as scaled scores. The summary statistics for scaled scores on each form are presented below in Table 42. 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 42 also includes the results for the SJT.

Table 42. Cognitive Scaled Score Test Statistics

Subtest	Form	Mean	*SD*	Min	Max	Reliability	*SEM*
VR	Form 1	602.95	79.87	300	880	0.76	39.13
	Form 2	603.01	79.26	300	900	0.76	38.83
	Form 3	606.38	83.53	300	900	0.78	39.18
	Form 4	599	80.01	300	900	0.76	39.2
	Form 5	600.34	79.71	300	900	0.76	39.05
DM	Form 1	627.97	84.95	300	890	0.78	39.85
	Form 2	620.52	83.75	300	900	0.78	39.28
	Form 3	634.18	91.34	300	900	0.81	39.81
	Form 4	628.08	89.41	300	900	0.81	38.97
	Form 5	627	80.87	300	890	0.77	38.78
QR	Form 1	663.21	110.53	300	900	0.82	46.89
	Form 2	663.15	109.88	300	900	0.82	46.62
	Form 3	673.49	114.1	300	900	0.81	49.74
	Form 4	654.88	107.45	300	900	0.82	45.59
	Form 5	647.08	105.21	300	900	0.82	44.64
Total Cognitive	Form 1	1,894.13	245.39	980	2,660	0.92	69.41
	Form 2	1,886.68	240.35	1,140	2,640	0.91	72.10
	Form 3	1,914.06	254.24	980	2,670	0.92	71.91
	Form 4	1,881.95	246.98	990	2,650	0.92	69.86
	Form 5	1,874.42	233.87	1,110	2,640	0.91	70.16
SJT	Form 1	604.26	73.71	300	754	0.82	31.27
	Form 2	602.53	72.84	300	768	0.83	30.03
	Form 3	599.03	69.99	300	774	0.82	29.69
	Form 4	598.81	76.38	300	767	0.82	32.41
	Form 5	597.43	74.24	300	749	0.83	30.61

6. Item Analysis

Each year, Pearson VUE undertakes item writing, pretesting, data analysis and statistical screening. New items are pretested along with operational items to establish their efficacy before being introduced into the operational item bank. At the end of each testing window, both operational and pretest items are analysed. The purpose of item analysis is to examine the item quality and determine whether items are suitable for future use.

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

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 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 43 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 43. Cognitive Items Passing the Quality Criteria

		VR		DM		QR	
		N	%	*N*	%	*N*	%
Operational	Pass	200	100%	155	100%	160	100%
	Fail	0	0%	0	0%	0	0%
	p < 10 or > 95	0	0%	0	0%	0	0%
	pBis <= 0.1	0	0%	0	0%	0	0%
	|*b*| >= 3	0	0%	0	0%	0	0%
Pretest	Pass	240	97%	234	98%	251	99%
	Fail	7	3%	4	2%	3	1%
	p < 10 or > 95	0	0%	0	0%	0	0%
	pBis <= 0.05	3	1%	4	2%	1	0%
	|*b*| >= 3	4	2%	0	0%	2	1%

In previous years, only a very small number of operational items failed the analysis each year. This year none of the operational items failed. For pretest items, some failures occurred in the subtests, mainly because of low discriminability. Notably, since 2022, there have been no pretest items for AR. As illustrated in Figure 25 and Figure 26, both operational and pretest item pass rates have continued to improve over time, resulting in excellent overall pass rates.

Figure 25. Proportion of Operational Items Failing Analysis 2017–2025

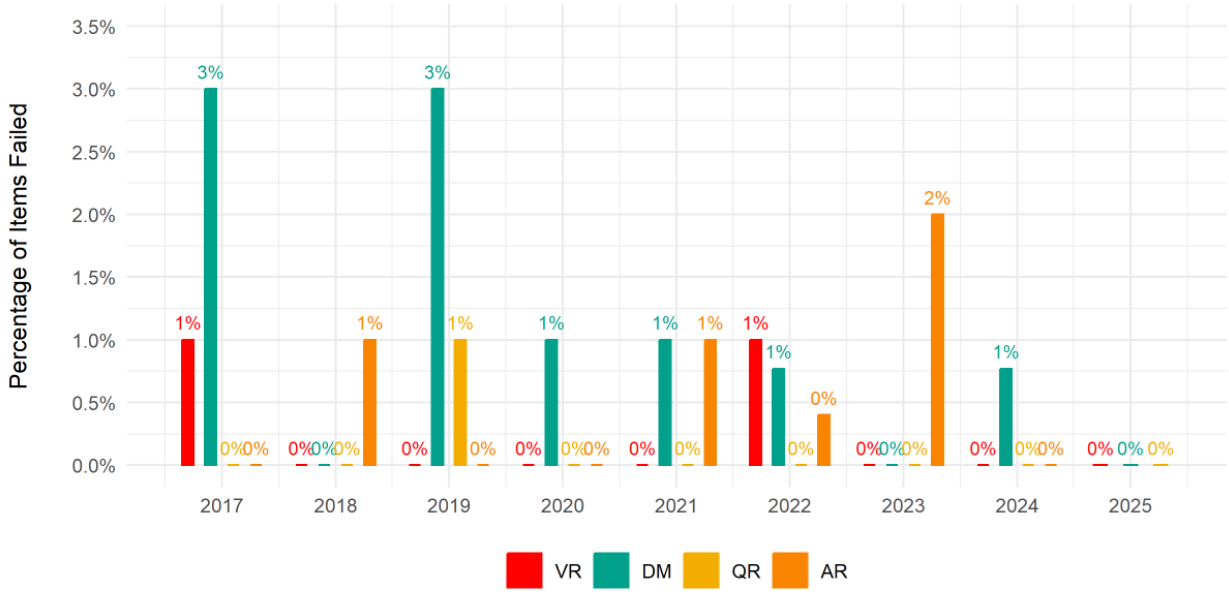


Figure 26. Proportion of Pretest Items Failing Analysis 2017–2025

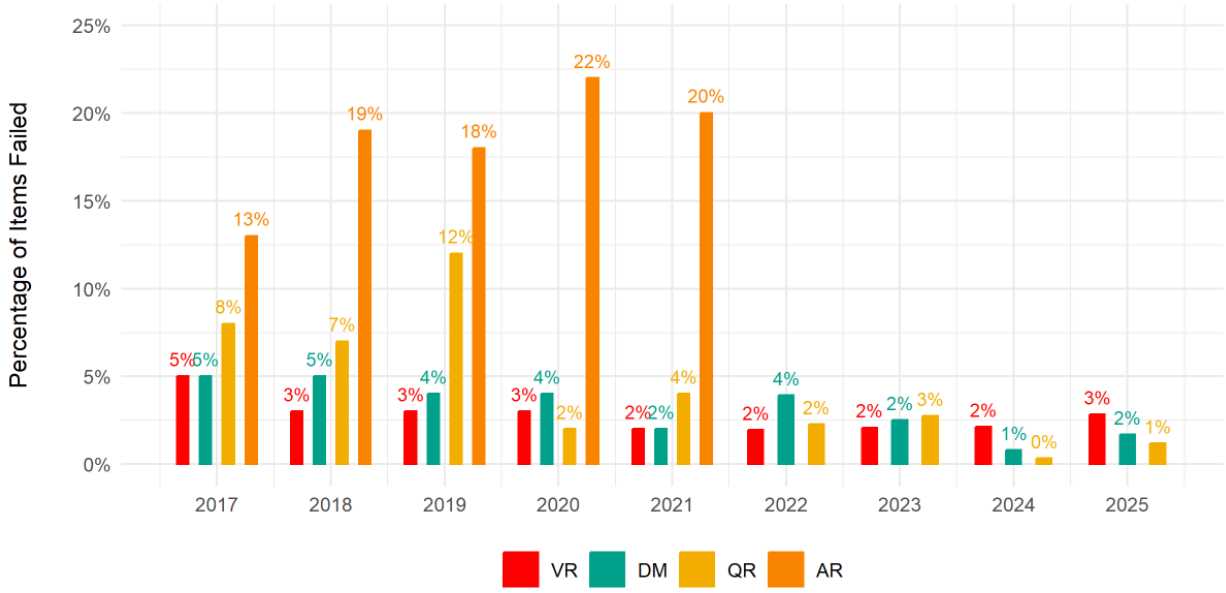


Table 44 shows a summary of the point biserial values. The maximum point biserial is 1, and higher values are better because they indicate that an item can discriminate well between strong and weak candidates. Given that the unscored items have not been tested before, it is expected that those items, on average, will discriminate less well than the scored items, and that is the case across all the cognitive subtests.

Table 44. Discrimination Summary Statistics

Scored/Unscored	Subtest	*N* Items	pBis			
			Mean	*SD*	Min	Max
Operational (Scored)	VR	200	0.30	0.06	0.15	0.43
	DM	155	0.36	0.10	0.13	0.64
	QR	160	0.45	0.06	0.28	0.63
Pretest (Unscored)	VR	247	0.28	0.09	-0.04	0.45
	DM	238	0.35	0.13	0.01	0.63
	QR	254	0.42	0.09	-0.02	0.64

Historically, the point biserial values for scored items have been high and stable, while those for unscored items have been slightly lower and less consistent, as shown in Figure 27. The point biserial for operational items has remained relatively stable, but pretest items have shown a noticeable increase in the past two years, with QR pretest items exhibiting a particularly large improvement. This indicates that the quality of pretest items has improved over time.

Figure 27. Point biserial 2017–2025

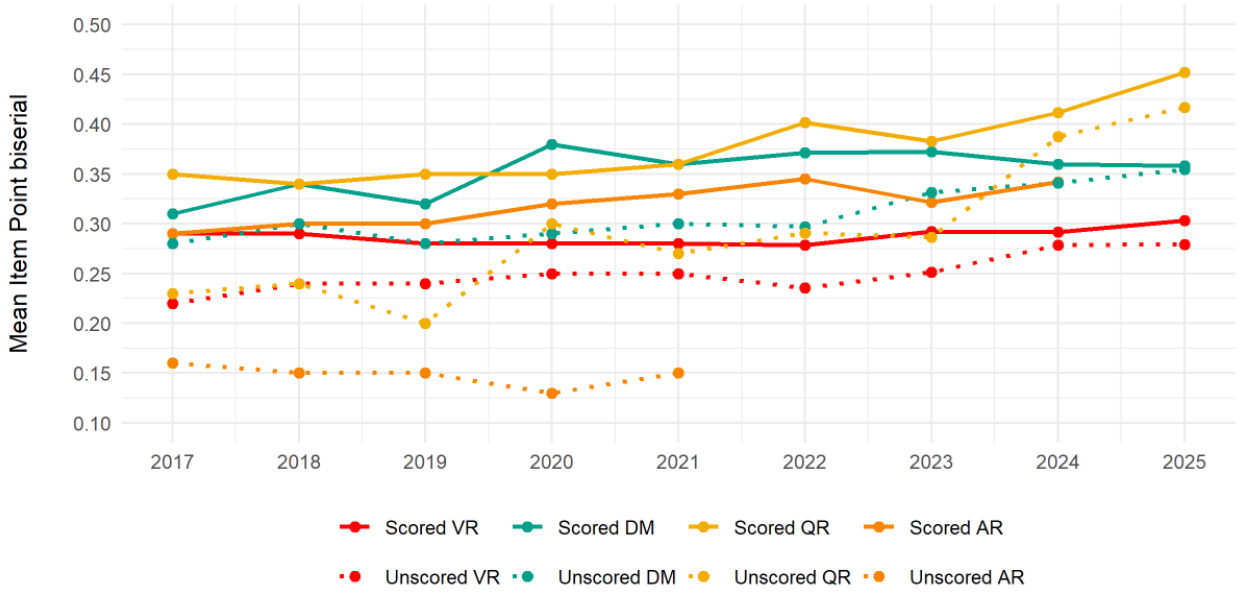


Table 45 presents a summary of the *p* values for the cognitive subtests. The *p* value indicates the proportion of candidates who answered each item correctly; higher *p* values mean the items are easier, while lower values indicate greater difficulty. Among the operational items, VR and DM were generally more challenging for candidates in 2025, whereas QR items tended to be easier. In the pretest pools, both DM and QR items appeared somewhat more difficult than those used operationally, but pretest VR items were actually easier compared to their operational counterparts.

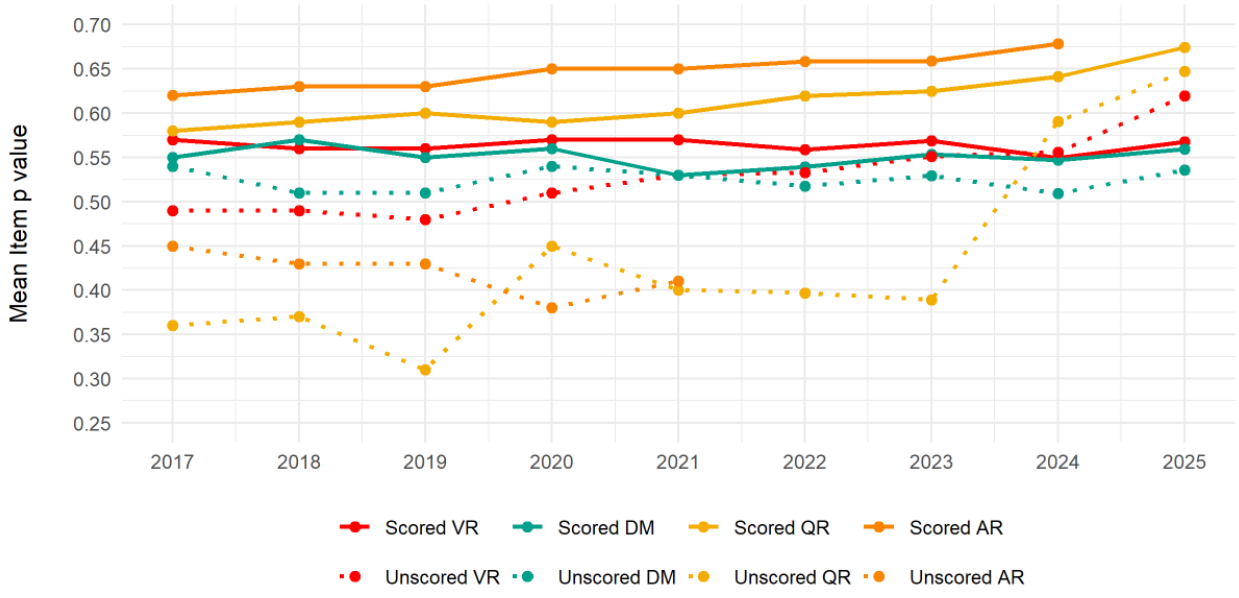
Table 45. p Value Summary Statistics

Scored/Unscored	Subtest	*N* Items	*p* Value			
			Mean	*SD*	Min	Max
Operational (Scored)	VR	200	0.57	0.14	0.24	0.85
	DM	155	0.56	0.15	0.18	0.92
	QR	160	0.67	0.13	0.36	0.89
Pretest (Unscored)	VR	247	0.62	0.18	0.19	0.97
	DM	238	0.54	0.18	0.12	0.95
	QR	254	0.65	0.18	0.19	0.96

Since 2017, pretesting has proven to be effective in identifying items that are either overly challenging or insufficiently demanding. Figure 28 demonstrates that items within the pretest pools generally exhibit higher difficulty, on average, compared to operational items. It is important to emphasise that subtests are equated annually, ensuring that fluctuations in individual item difficulty do not impact the ability level required for candidates to attain a specific scaled score. QR pretest items displayed a marked increase in *p* value last year, with a further slight increase this year, resulting in greater similarity to operational items. This indicates that the QR items written for 2025 continue to be more discriminative and easier, consistent with the trend observed in 2024. The progression reflects a sustained effort to align pretest item difficulty with that of the operational bank

as needed. Notably, the gap between QR pretest items and the operational bank has narrowed substantially. Overall, all subtests are achieving closer alignment between pretest item difficulty and the operational bank. While VR items were highly aligned in 2024 and have diverged slightly this year, they remain relatively close; however, unlike other subtests, this year's VR pretest items appear to be easier than their operational counterparts.

Figure 28. p Value 2017–2025



The VR subtest consists of four-option multiple-choice items and three-option true/false/can't tell items. Table 46 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 biserial in the pretest pool shows that pretesting is successfully removing items that do not discriminate effectively.

Table 46. 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	160	0.31	0.05	0.58	0.14
	True/False/Can't Tell	40	0.26	0.05	0.54	0.13
Pretest (Unscored)	Multiple Choice	215	0.29	0.08	0.61	0.18
	True/False/Can't Tell	32	0.23	0.10	0.66	0.16

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 47.

Table 47. 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	55	0.43	0.10	0.48	0.15
	Multiple Choice	100	0.32	0.08	0.60	0.14
Pretest (Unscored)	Drag and Drop	97	0.45	0.10	0.51	0.16
	Multiple Choice	141	0.29	0.11	0.55	0.19

The DM subtest is composed of items that vary not only in response format but also in item type. Within the drag-and-drop category, there is a distinction in difficulty and discriminative ability between item types. Specifically, interpreting information items have been found to be more challenging for test-takers than syllogism items. However, syllogism items, while less difficult, show a slightly higher ability to discriminate between different levels of performance. These findings are summarised in Table 48.

For the multiple-choice items, the statistical reasoning and Venn diagram items showed the greatest discriminative power among all item types. In 2024, statistical reasoning items were found to be the most difficult in the DM subtest. In contrast, this year, the logical puzzles category has emerged as the most difficult item type. Meanwhile, Venn diagram items remain the easiest, a trend that is consistent with the previous year's results.

Table 48. 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	25	0.40	0.10	0.45	0.14
		Syllogisms	30	0.46	0.09	0.51	0.15
	Multiple Choice	Logical Puzzles	25	0.25	0.04	0.52	0.10
		Statistical Reasoning	20	0.36	0.08	0.55	0.14
		Assumptions Recognition	20	0.26	0.06	0.62	0.14
		Venn Diagrams	35	0.37	0.06	0.68	0.13
Pretest (Unscored)	Drag and Drop	Information Interpretation	49	0.44	0.10	0.46	0.15
		Syllogisms	48	0.46	0.10	0.56	0.15
	Multiple Choice	Logical Puzzles	32	0.21	0.08	0.54	0.19
		Statistical Reasoning	40	0.32	0.11	0.49	0.17
		Assumptions Recognition	19	0.24	0.10	0.60	0.15
		Venn Diagrams	50	0.33	0.09	0.60	0.21

The QR subtest consists of both item sets and standalone items. Each item set is composed of four related items. As with the pretest pool as a whole, the pretest items discriminate slightly less well on average than the ones that have already

been pretested prior to appearing in the 2025 exam, as shown in Table 49. Furthermore, items that are part of an item set showed a slightly greater discriminative ability than standalone items. Despite this difference in discrimination, both item types are found to be similar in terms of their operational difficulties.

Table 49. 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	140	0.46	0.06	0.68	0.13
	Standalone	20	0.42	0.08	0.66	0.14
Pretest (Unscored)	Item Set	234	0.42	0.09	0.64	0.18
	Standalone	20	0.40	0.09	0.76	0.16

Item Analysis for SEN

An additional analysis was performed to examine whether the items perform differently for exams with accommodations. Overall, the item performances did not show substantial differences between the two sets 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 50. 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 50. Item Analysis of UCAT and UCATSEN

Scored/Unscored	Subtest	Statistics	UCAT		UCATSEN	
			Mean	*SD*	Mean	*SD*
Operational (Scored)	VR	*p* Value	0.56	0.14	0.60	0.14
		Point biserial	0.30	0.06	0.30	0.07
		IRT *b*	-0.20	0.65	-0.21	0.68
	DM	Facility	0.73	0.27	0.79	0.28
		Point biserial	0.36	0.11	0.33	0.09
		IRT *b*	0.24	0.72	0.20	0.82
	QR	*p* Value	0.67	0.13	0.73	0.13
		Point biserial	0.45	0.06	0.43	0.07
		IRT *b*	-0.32	0.72	-0.34	0.80
Pretest (Unscored)	VR	*p* Value	0.62	0.18	0.66	0.19
		Point biserial	0.28	0.09	0.26	0.20
		IRT *b*	-0.51	0.97	-0.62	1.15
	DM	Facility	0.74	0.34	0.79	0.37
		Point biserial	0.35	0.13	0.36	0.21
		IRT *b*	0.34	0.89	0.35	1.05

Scored/Unscored	Subtest	Statistics	UCAT		UCATSEN	
			Mean	*SD*	Mean	*SD*
	QR	*p* Value	0.64	0.18	0.69	0.19
		Point biserial	0.42	0.09	0.38	0.20
		IRT *b*	-0.24	1.13	-0.20	1.34

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 2025. This section presents the performance of test items across the UK and ANZ population of the 2025 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 51 compares the summary statistics for the operational item analysis of the UCAT 2025 and the UCAT ANZ 2025. 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-dependent, 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 51 shows that for VR, the 2025 UCAT and UCAT ANZ had very similar mean IRT difficulty values, while DM and QR have slightly different mean IRT b values.

Table 51. Comparison of Operational Item Statistics: UCAT & UCAT ANZ 2025

Subtest	Item Statistics	*N* Items	UCAT		UCAT ANZ	
			Mean	*SD*	Mean	*SD*
VR	*p* Value	200	0.57	0.14	0.60	0.14
	Point biserial	200	0.30	0.06	0.32	0.06
	IRT Difficulty	200	-0.20	0.65	-0.20	0.66
DM	Facility	155	0.56	0.15	0.59	0.15
	Point biserial	155	0.36	0.10	0.38	0.11
	IRT Difficulty	155	0.24	0.72	0.20	0.77
QR	*p* Value	160	0.67	0.13	0.70	0.11
	Point biserial	160	0.45	0.06	0.48	0.06
	IRT Difficulty	160	-0.33	0.73	-0.28	0.69

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 52 summarises the number of items showing drift in the UCAT since 2017 and Table 53 in the UCAT ANZ since 2019. The difficulty of the items is anchored on historical UCAT data, primarily based on UK candidates. If there are significantly more or fewer drifted items in UCAT ANZ compared to UCAT, this may indicate that regional differences influence how content is perceived, impacting observed item difficulty. In 2025, the number of drift items in UCAT and UCAT ANZ is comparable, suggesting that both cohorts are interacting with item content in similar ways. The Content Team reviewed items with different drift patterns but found no clear explanation related to cultural sensitivity.

Table 52. Number of Operational Items Showing Drift in UCAT

Subtest	UCAT								
	2017	2018	2019	2020	2021	2022	2023	2024	2025
VR	2 (2%)	3 (3%)	6 (3%)	4 (2%)	4 (2%)	5 (3%)	6 (4%)	4 (2%)	3 (2%)
DM	11 (14%)	6 (8%)	17 (13%)	37 (28%)	12 (9%)	7 (5%)	3 (3%)	12 (9%)	13 (8%)
QR	2 (2%)	0 (0%)	1 (1%)	0 (0%)	2 (1%)	6 (4%)	2 (2%)	9 (6%)	13 (8%)
AR	7 (5%)	5 (3%)	21 (8%)	25 (10%)	40 (16%)	19 (8%)	5 (3%)	19 (8%)	-

Table 53. Number of Operational Items Showing Drift in UCAT ANZ

Subtest	UCAT ANZ						
	2019	2020	2021	2022	2023	2024	2025
VR	12 (10%)	13 (6%)	13 (6%)	8 (4%)	9 (6%)	4 (2%)	8 (4%)
DM	7 (9%)	47 (36%)	11 (8%)	9 (7%)	8 (8%)	10 (8%)	19 (12%)
QR	3 (3%)	2 (1%)	4 (2%)	5 (3%)	4 (3%)	11 (7%)	9 (6%)
AR	22 (15%)	24 (10%)	37 (15%)	13 (5%)	7 (4%)	19 (8%)	-

At present, it is recommended that the degree of drift be monitored in 2026. We would not recommend taking any action to create a separate item bank for the UCAT ANZ at this time.

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 54. 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 54. Candidate Removal Summary for SJT Item Analysis

Statistic	Criteria	Number of Candidates Removed
1. *Z* score of the scaled score	*Z* score < -4.222	0
2. High number of missing responses	> 1 blank response on operational items	1,344
3. Low completion time	Drop in score based on response time	0

The following item statistics are calculated for the SJT items:

- Item facility: the mean score on the items as a percentage of the maximum score available. It represents the difficulty of the item.
- Item *SD*: the *SD* of the scores on the items. It gives an indication of how well the item is differentiating among candidates.
- Item partial correlation: the correlation of the item score with the total score for the operational items and the scaled score for the pretest items. It compares how individuals perform on a given item with how they perform on the test overall and is a measure of discrimination. Item correlations can be interpreted in the following way:
    - Below 0.1 – poor correlation with the test overall, and items within this band are unlikely to be used in an operational test.
    - 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 > 0.10

Since 2023, the quality criteria for SJT items have been adjusted to align with those used for cognitive items. Specifically, the item partial criterion has been reduced from 0.13 to 0.10, and item partial and facility is now rounded to a whole number prior to the application of the criterion. The new criteria are slightly more lenient than the previous ones, allowing a slightly higher number of operational and pretest items to be classified as successful. These adjustments are intended to support the ongoing development and refinement of the item bank, and consistency with the criteria used for the cognitive tests. Table 55 shows the number of items that met and did not meet the quality criteria.

Table 55. SJT Item Quality Criteria

	Type	Criteria	All		Appropriateness		Direct Speech		Importance	
			N	%	*N*	%	*N*	%	*N*	%
Operational	Rating Items	Met	171	86%	64	82%	30	91%	77	89%
		Not met	27	14%	14	18%	3	9%	10	11%
	Most/Least Items	Met	9	100 %						
		Not met	0	0%						
	Type	**Criteria**	**All**		**Appropriateness & Appropriateness (Speech)** Note				**Importance**	
			N	%	*N*		%		*N*	%
Pretest	Rating Items	Met	180	51%	119		51%		61	50%
		Not met	173	49%	113		49%		60	50%
	Most/Least Items	Met	14	78%						
		Not met	4	22%						

Note. In 2025, the item category Appropriateness and Direct Speech are updated to Appropriateness and Appropriateness (Speech) for new pilot items and are grouped into one category.

The proportion of items meeting quality criteria has improved slightly, as shown in Table 55 and Figure 29. Pretest most/least items not meeting the criteria decreased from 37% in 2023 to 18% in 2024, then rose to 22% in 2025; though, given the small number of items, a difference of one item can look relatively large proportionately. For standard rating items, the percentage not meeting criteria dropped from 57% in 2022 and 2023 to 46% in 2024, but increased to 49% in 2025. However, the operational items not meeting the criteria decreased in 2025. These results show ongoing efforts to improve pilot item writing, though some year-to-year fluctuations remain.

Figure 29. Proportion of SJT Items Failing Analysis 2017–2025

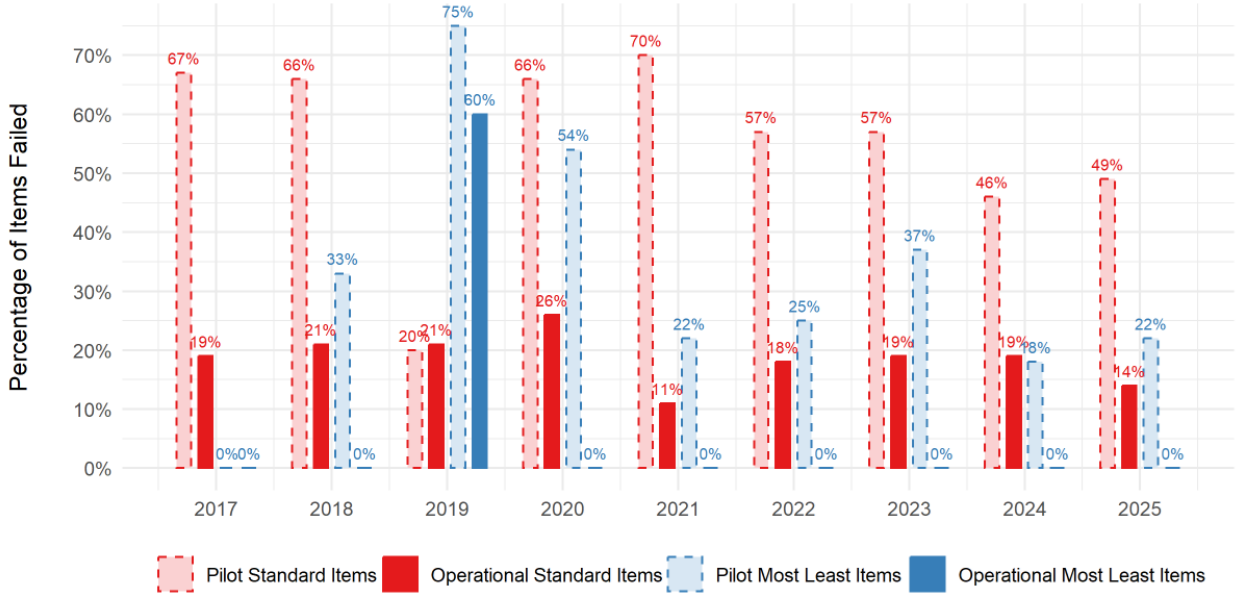


Table 56 provides a summary of the item analysis results of operational SJT items. This table presents key statistical measures for the performance of SJT items that have been administered operationally.

Table 56. Operational SJT Item Analysis Summary

	Mean	*SD*	Min	Max
Item Mean	2.89	1.14	0.62	7.37
Item *SD*	1.04	0.33	0.25	2.75
Item Partial Correlation	0.23	0.10	-0.04	0.49
Item Total Facility	0.73	0.18	0.21	0.97

Since 2017, both the item mean score and facility have generally shown an upward trend, as demonstrated in Figure 30, suggesting that test items have become somewhat easier over time. In response, targeted measures were implemented to increase item difficulty and achieve a balanced assessment. Consequently, item facility peaked in 2023, followed by a gradual decrease in 2024 and 2025. Notably, the SJT facility for the current year is slightly lower than last year's figure, returning to levels comparable to those observed in 2021, which has been intentional as part of test construction.

Figure 30. Average Item Facility of Operational SJT Items 2017–2025

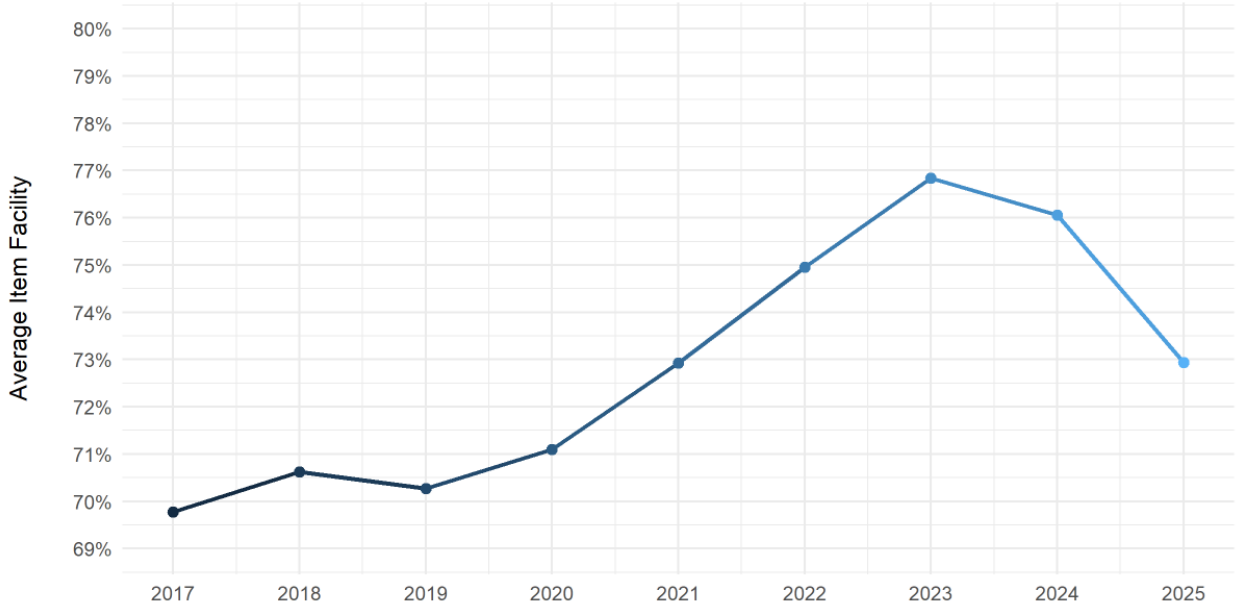


Figure 31 also indicates a decline in item partial correlation, suggesting that although the test was more challenging, its capacity to differentiate between higher-performing and lower-performing candidates decreased. This implies that the items were generally less effective at distinguishing candidate performance compared to those in 2022 and 2023. Nevertheless, the item partial correlation remains within the expected range. In line with the item facility, the item partial correlation has returned to levels observed in 2021, indicating that the operational quality and specification of SJT items in 2025 closely resemble those in 2021.

Figure 31. Average Item Partial Correlation of Operational SJT Items 2017–2025

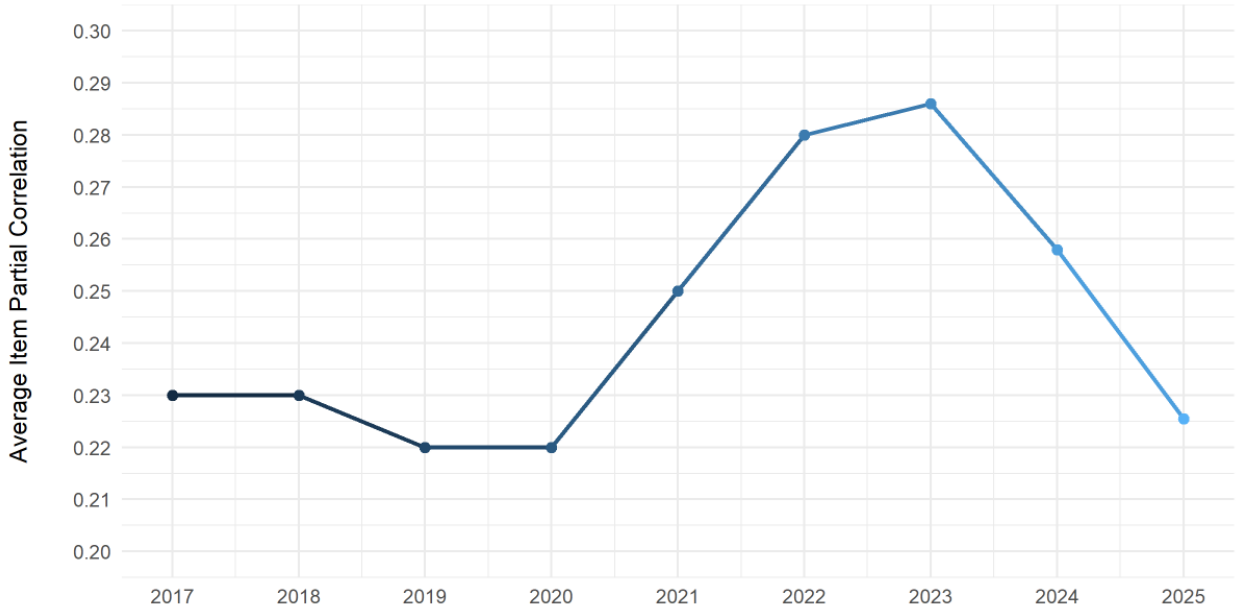


Table 57 summarises the statistics for the SJT pretest items. While the most/least items demonstrated slightly higher discriminating ability compared to the standard rating items, they also showed a relatively higher average item total facility.

Table 57. SJT Pretest Item Summary Statistics

	Statistic	Item Mean	Item *SD*	Item Partial	Item Total Facility
Rating Items	Mean	2.83	0.84	0.14	0.76
	SD	0.99	0.29	0.10	0.21
	Min	0.47	0.14	-0.08	0.16
	Max	3.99	1.56	0.46	1.00
Most/Least	Mean	7.20	1.31	0.18	0.90
	SD	0.45	0.38	0.10	0.06
	Min	6.06	0.90	-0.08	0.76
	Max	7.72	2.01	0.32	0.97

Differential Item Functioning (DIF)

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.

Method of DIF Detection

For the 2025 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.

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.

DIF Results

The DIF results are reported below for each demographic group. Table 58 shows DIF in relation to gender. One pretest VR item was found to exhibit Category C DIF favouring Female over Male.

Table 58. Gender DIF

Group	Code	VR		DM		QR		SJT	
		N	%	*N*	%	*N*	%	*N*	%
Operational	A	198	99%	153	99%	160	100%	204	99%
	B	2	1%	2	1%	0	0%	3	1%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
Pretest	A	245	99%	237	100%	254	100%	191	51%
	B	1	0%	1	0%	0	0%	9	2%
	C	1	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	171	46%

Since 2022, the criteria for age comparisons have been revised to allow for more items to be evaluated. The current comparison is between candidates under 20 years of age and those over 25, rather than the previous comparison of under 20 and over 35, as outlined in Table 59. Previously, the limited number of candidates over 35 made meaningful comparisons challenging. With the updated groups, more DIF has been identified in recent years, as prior analyses were unable to detect potential biases within the new comparison group, and these items were not routinely removed from the item bank.

Two operational VR items were identified with Category C DIF, favouring older candidates. In DM, fifteen Category C DIF items were identified, with two

favouring older candidates and thirteen favouring younger candidates. No QR items exhibited age Category C DIF.

Similar to 2024, the relatively larger number of operational items identified with DIF, but not pretest items, is likely a result of the updated age comparison grouping introduced in 2022. It is uncommon to see a large number of operational items show DIF, as these items had already passed DIF evaluation before being added to the operational bank. However, in this case, the increase is understandable due to the change in grouping. Previously, these items may not have been adequately assessed due to the smaller number of candidates older than 35 and the differences in characteristics of candidates in this age group. This increase in operational DIF items demonstrates that the updated comparison criteria have been effective in identifying items that may have shown bias but were previously unidentified. This adjustment has contributed to improving the item bank and reducing bias across the test overall.

Table 59. Age DIF

Group	Code	VR		DM		QR		SJT	
		N	%	*N*	%	*N*	%	*N*	%
Operational	A	191	96%	125	81%	154	96%	205	99%
	B	7	4%	15	10%	6	4%	2	1%
	C	2	1%	15	10%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
Pretest	A	0	0%	10	4%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	247	100%	228	96%	254	100%	371	100%

For ethnicity, there are typically enough items to reliably categorise DIF for operational items. However, many pretest comparisons are not feasible due to low candidate numbers, as pretest items involve smaller sample sizes. It is also important to note that the ethnicity question options have changed since 2022, with the “UK - Chinese” category no longer listed separately. Additionally, since 2022, a comparison between White and Non-White candidates has been included.

Table 60 identifies five instances of Category C DIF in the ethnicity comparisons for operational items; three instances are linked to the same item in DM, two items were in QR, and the remaining one was in SJT. The DM item showing bias favoured White candidates over Black and Asian candidates and overall favoured White candidates over Non-White candidates. One QR item favoured White candidates over Black candidates, and the other one favoured Black candidates over White candidates. The SJT item favoured White candidates over Black candidates. Additionally, two pretest items were found to be exhibiting Category C DIF. A VR pretest item was found to be favouring White candidates over Asian candidates, and a DM pretest item was found to be favouring Black candidates over White candidates.

Table 60. Ethnicity DIF

Type	Group	Code	VR		DM		QR		SJT	
			N	%	N	%	N	%	N	%
Operational	White/ Black	A	199	100%	146	94%	150	94%	200	97%
		B	1	0%	8	5%	8	5%	6	3%
		C	0	0%	1	1%	2	1%	1	0%
		NA	0	0%	0	0%	0	0%	0	0%
	White/ Asian	A	200	100%	151	97%	157	98%	201	97%
		B	0	0%	3	2%	3	2%	6	3%
		C	0	0%	1	1%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%
	White/ Mixed	A	200	100%	154	99%	160	100%	207	100 %
		B	0	0%	1	1%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%
	White/ Non-White	A	200	100%	153	99%	158	99%	205	99%
		B	0	0%	1	1%	2	1%	2	1%
		C	0	0%	1	1%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%
Pretest	White/ Black	A	147	60%	119	50%	106	42%	8	2%
		B	0	0%	0	0%	0	0%	4	1%
		C	0	0%	1	0%	0	0%	0	0%
		NA	100	40%	118	50%	148	58%	359	97%
	White/ Asian	A	246	100%	235	99%	254	100%	16	4%
		B	0	0%	1	0%	0	0%	2	1%
		C	1	0%	0	0%	0	0%	0	0%
		NA	0	0%	2	1%	0	0%	353	95%
	White/ Mixed	A	0	0%	2	1%	0	0%	0	0%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	247	100%	236	99%	254	100%	371	100 %
	White/ Non-White	A	247	100%	238	100%	254	100%	18	5%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	353	95%

Since 2022, comparisons between SEC1 and non-SEC1 candidates have been incorporated to facilitate more comprehensive analyses. This year, as indicated in Table 61, no items were identified in the Category C DIF for SEC.

Table 61. SEC DIF

Type	Group	Code	VR		DM		QR		SJT	
			N	%	*N*	%	*N*	%	*N*	%
Operational	SEC 1/2	A	200	100%	155	100%	160	100%	108	52%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	99	48%
	SEC 1/3	A	200	100%	155	100%	160	100%	207	100%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%
	SEC1/4	A	199	100%	155	100%	159	99%	184	89%
		B	1	0%	0	0%	1	1%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	23	11%
	SEC 1/5	A	200	100%	155	100%	160	100%	207	100%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%
	SEC 1/(2-5)	A	200	100%	155	100%	160	100%	207	100%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	0	0%	0	0%	0	0%
Pretest	SEC 1/2	A	0	0%	0	0%	0	0%	0	0%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	247	100%	237	100%	254	100%	371	100%
	SEC 1/3	A	175	71%	128	54%	160	63%	1	0%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	72	29%	110	46%	94	37%	370	100%
	SEC 1/4	A	0	0%	0	0%	0	0%	0	0%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	247	100%	238	100%	254	100%	371	100%
	SEC 1/5	A	15	6%	73	31%	3	1%	0	0%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	232	94%	165	69%	251	99%	371	100%
	SEC 1/(2-5)	A	247	100%	237	100%	254	100%	18	5%
		B	0	0%	0	0%	0	0%	0	0%
		C	0	0%	0	0%	0	0%	0	0%
		NA	0	0%	1	0%	0	0%	353	95%

As shown in Table 62, two Category C DIF items were identified in the comparison between candidates with an honours degree or higher and those without. Both items were DM items, one pretest and one operational (one QR and one DM), and both favoured candidates without degree-level education over those that had degree-level education.

Table 62. Honours Degree DIF

Type	Code	VR		DM		QR		SJT	
		N	%	*N*	%	*N*	%	*N*	%
Operational	A	200	100%	145	94%	160	100%	206	100%
	B	0	0%	9	6%	0	0%	1	0%
	C	0	0%	1	1%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
Pretest	A	247	100%	229	96%	254	100%	18	5%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	1	0%	0	0%	0	0%
	NA	0	0%	8	3%	0	0%	353	95%

Table 63 presents the comparison between candidates who reported English as their first or primary language and those who did not. No items were identified as Category C DIF for this language comparison.

Table 63. English as First Language DIF

Group	Code	VR		DM		QR		SJT	
		N	%	*N*	%	*N*	%	*N*	%
Operational	A	199	100%	155	100%	158	99%	207	100%
	B	1	0%	0	0%	2	1%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
Pretest	A	247	100%	237	100%	254	100%	28	8%
	B	0	0%	1	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	343	92%

As shown in Table 64, one Category C DIF item was identified in the comparison between candidates who reported the UK as their residence and those who did not. A pretest DM item was found to favour non-UK residents over UK residents.

Table 64. Residency DIF

Group	Code	VR		DM		QR		AR		SJT	
		N	%	N	%	N	%	N	%	N	%
Operational	A	199	100%	151	97%	157	98%	206	100%	199	100%
	B	1	0%	4	3%	3	2%	1	0%	1	0%
	C	0	0%	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%	0	0%
Pretest	A	247	100%	237	100%	254	100%	23	6%	247	100%
	B	0	0%	0	0%	0	0%	0	0%	0	0%
	C	0	0%	1	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	348	94%	0	0%

In summary, 26 Category C DIF items were found in 2025, consisting of 21 operational and 5 pretest items. This count is close to that in 2024, which saw 35 Category C DIF items (with 21 operational and 14 pretest items), and 24 in 2023 (8

operational and 16 pretest items). These numbers mark notable increases from previous years: 10 items in 2022 and 13 in 2021. The rise can be linked to significant changes in test structure in 2025, a growth in international partner universities since 2023, and adjustments to age DIF and ethnicity comparison groups from 2022 onwards.

To promote fairness, all identified DIF items have been removed from the item bank and will not appear in future tests. Continued efforts will include reviewing these items and refining item development procedures to further reduce potential bias.

7. Summary

In 2025, UCAT underwent a major restructure, including the removal of AR, the lengthening of DM, additional time for VR and QR with corresponding scaling adjustments, and the SJT banding setting approach was updated. The test performed as intended, with the mean scaled scores of all subtests within reasonable deviation from 2024 results after the implemented changes. The speededness of the subtests was slightly alleviated following the additional allocated time. Updates to the SJT band setting approach also resulted in the distribution aligning more closely with targets.

The composition of candidates in 2025 remains largely consistent with that observed in the previous year. However, a significant change this year is the record number of candidates sitting the exam. This continued expansion of the exam is demonstrated by a notable increase in candidate volume, with 2025 registering nearly 10% more candidates compared to 2024. Furthermore, when compared to candidate volume from 2017, there is an increase of more than 65%, underscoring the substantial growth in candidate numbers over recent years.

The item analysis for this year demonstrates continued improvement in the quality of items within the bank. Notably, all operational items in the cognitive subtests met the required statistical criteria, which reflects a gradual enhancement in item quality over time. This achievement highlights the rigorous development and review processes in place for operational items.

Consistent with trends observed in previous years, the passing rates for pretest items within the cognitive subtests remained exceptionally high. This suggests that the pretest item pool continues to meet the expected standards for candidate performance and item validity.

Similarly, the passing rates for the SJT subtest were largely in line with previous years' results. A gradual increase in the difficulty of SJT items has been noted,

continuing the trend established in 2024 and reflecting the overall direction of test development. However, it has been noted that there has been a corresponding decrease in item discriminability. This relationship between increasing difficulty and declining discriminability warrants further investigation. Efforts will focus on identifying strategies to enhance item difficulty while maintaining robust discriminability to ensure the continued effectiveness and fairness of the SJT subtest, and, this will be considered alongside the investigation of the use of IRT for this subtest.

Consistent with the findings from 2024, the DIF analysis in 2025 indicated that only a very small proportion of test items were classified as Category C, which identifies items exhibiting significant differential item functioning. However, the number of Category C items observed in 2025 was slightly higher than those reported in earlier years, such as 2021. Several factors may have contributed to this modest increase. Key influences include the various structural changes made to the test in recent years, adjustments in the categorisation of DIF, and a growing population of international candidates. These elements combined are likely to have impacted the identification and distribution of Category C items, reflecting the evolving nature of the test and its candidate demographics.

Candidates who require special accommodations continue to make up a very small percentage of the total candidate population. Among those who receive accommodations, the UCATSEN group constitutes the largest segment. With the removal of the AR subtest and the subsequent decrease in the total cognitive scaled score, it has become challenging to directly compare total score differences between the standard and extended versions of the exam based on historic data. Nonetheless, the observed degree of difference between the two exam formats appears to remain broadly consistent.

While several substantive changes were introduced, the results outside these modifications remained stable, reflecting the ongoing integrity of the assessment process. The demographic composition of candidates sitting the exam in 2025 was largely unchanged in comparison to earlier years, with differences between demographic groups maintaining a consistent pattern over time. The quality of the test forms used in 2025 was upheld through rigorous design and analysis. Measurement error was kept appropriately low, ensuring reliable results for candidates. Furthermore, the different test forms were well-balanced, as evidenced by the consistency in average scores across all versions of the exam administered during the year.

Recommendations

Given the substantial changes implemented this year, it is advisable to maintain the test format largely consistent with the current structure. This approach will help preserve the stability of the assessment and provide an opportunity to closely monitor the effects of the recent modifications over time.

It is recommended to further rescale the QR subtest. Specifically, scaling down the QR subtest by an additional 10 scale score points is proposed. This

suggestion arises from the observation that the mean QR score continues to demonstrate an upward trend, even after it was previously scaled down by 10 points in the current year. Further adjustment aims to better align the QR subtest scores with the intended distribution and maintain the overall balance among subtests.

Ongoing research is being conducted to evaluate the potential application of IRT in the SJT subtest. The objective of this investigation is to achieve a more even distribution of SJT scores. In addition, further analysis has been performed on factors to be considered during item writing and form assembly. These investigations are intended to improve the quality of new items and ensure effective form balancing. Areas under review include the use of text analysis and considerations of additional factors, such as the recency of item usage, to enhance both item quality and the balance of test forms.

8. References

Bala, L., Pedder, S., Sam, A., & Brown, C. (2022). Assessing the predictive validity of the UCAT—A systematic review and narrative synthesis. *Medical Teacher, 44*(4), 401-409. doi:10.1080/0142159X.2021.1998401

Greatrix, R., Nicholson, S., & Anderson, S. (2021). Does the UKCAT predict performance in medical and dental school? A systematic review. *BMJ Open, 11*(1), e040128. doi:10.1136/bmjopen-2020-040128

Paton, L. W., & Tiffin, P. A. (2024). *Exploring performance differences between UCAT candidates who sit standard and extended versions of the test: report for the UCAT Board.* UCAT.

Paton, L., McManus, I., Cheung, K., Smith, D., & Tiffin, P. (2022). Can achievement at medical admission tests predict future performance in postgraduate clinical assessments? A UK-based national cohort study. *BMJ Open, 12*(2), e056129. doi:10.1136/bmjopen-2021-056129

Tiffin, P., Mwandigha, L., Paton, L., Hesselgreaves, H., McLachlan, J., Finn, G., & Kasim, A. (2016). Predictive validity of the UKCAT for medical school undergraduate performance: A national prospective cohort study. *BMC Medicine, 14*(1), 1-13. doi:10.1186/s12916-016-0682-7