# University Clinical Aptitude Test (UCAT)

**Technical Report**
**Testing Interval: 8 July 2024 to 26 September 2024**

**Prepared by:**
Pearson VUE
30 April 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

# 1. Executive Summary

The University Clinical Aptitude Test (UCAT) was administered in 2024 from 8 July 2024 to 26 September 2024. This report covers the 37,913 exams delivered during that period, marking a notable increase (6.4%) compared to 2023. The exam was delivered in two modes: online and test centre. Online test delivery accounted for less than 1% of candidates, making it unreliable to compare results between these two groups.

This report includes results from seven UCAT versions designed for candidates with special educational needs (SEN). SEN candidates represented a small proportion of test-takers, with UCATSEN being the most frequently used accommodation. As in previous years, candidates using SEN versions outperformed those taking the non-SEN version. Following the implementation of restrictions on the Pause-the-clock feature, there is no longer evidence of misuse. Usage patterns demonstrate that the feature effectively addresses the diverse needs of candidates, as shown by its varied applications. Additionally, the observation that eligible candidates often do not utilise the full time allowed suggests that the current accommodations are adequate. Therefore, no further adjustments to this feature are recommended.

Each UCAT consists of five subtests. In 2024, further rescaling adjustments were applied to better balance scaled scores across subtests. This resulted in a higher mean scaled score for Verbal Reasoning (VR) and lower scores for Quantitative Reasoning (QR) and Abstract Reasoning (AR), bringing the subtest averages closer together. Without the rescaling, the mean scaled scores for VR, QR, Decision Making (DM), and AR would have shown small deviations, with a slight decrease in VR and slight increases in QR and AR, partially counteracting the effect of the rescaling. However, these changes remained largely in line with score expectations. The Situational Judgement Test (SJT) banding distribution, on the other hand, showed fewer candidates in higher bands and more in lower bands, reflecting overall lower performance than anticipated.

The 2024 UCAT consisted of five test forms. The reliabilities of the forms were good, and the corresponding standard errors of measurement (*SEM*s) were low and consistent with previous years. Efforts to balance the difficulty and performance of test forms resulted in consistent average scores across forms.

The cognitive subtests remained speeded to a degree. Most candidates used all available time, and average time usage was close to the limit. Speededness was notably reduced in QR and AR, enabling better performance, while VR became slightly more speeded due to item ratio adjustments. Speededness was lowest in the SEN versions, where candidates had additional time. The SJT subtest remained the least speeded section overall.

Demographic trends in 2024 largely mirrored those of past years. The continued growth of international partner universities contributed to a record number of candidates, with non-UK candidates representing a larger proportion of the cohort. UK-Asian candidates remained the largest ethnic group among UK-based test-takers, while the proportion of UK-White candidates declined further. Candidates with a higher socio-economic classification (SEC), those of white ethnicity, UK residents, and English as a first language speakers continued to achieve higher scores across all subtests. Male candidates outperformed female candidates in the cognitive subtests, while female candidates scored higher in the SJT.

Item analysis for cognitive subtests showed satisfactory quality for most operational items, with improvements in passing rates for both operational and pretest items. In QR, pretest items were easier and more discriminating, addressing imbalances in the item bank. SJT item passing rates improved, partly due to more lenient criteria introduced this year matching that of the cognitive tests. An increase in Category C DIF items was observed in operational items but not in pretest items, suggesting that the rise was driven by changes in grouping criteria and the candidate sample rather than item quality.

In conclusion, the results of the 2024 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 in candidate composition and rescaling adjustments, the UCAT continues to provide a reliable measure of cognitive and non-cognitive abilities.

# 2. 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 a good clinician. It is not an exam that measures student achievement and therefore it does not contain any curriculum or science content.

This report covers the 2024 UCAT that was delivered from 8 July 2024 to 26 September 2024. As outlined in Section 3, the exam consisted of five subtests that each contained between 29 and 69 items. The design of the exam remained unchanged from the previous year. However, as in 2023, the scaling of three subtests was adjusted. The VR subtest was scaled up by 20 points, while the QR and AR subtests were each scaled down by 10 points.

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, usage of the pause-the-clock feature offered for the first time in 2023, 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.

# 3. Exam Design 2024

The 2024 UCAT consisted of five balanced test forms, each with five 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 | 21 minutes allowed on items and 1 minute for instruction |
| DM | 1 testlet of 26 items | 3 items | 29 | 31 minutes allowed on items and 1 minute for instruction |
| QR | 8 testlets of 4 items | 1 testlet of 4 items | 36 | 25 minutes allowed on items and 1 minute for instruction |
| AR | 10 testlets of 5 items | 0 items | 50 | 12 minutes allowed on items and 1 minute for instruction |
| SJT | 20 testlets of 1 to 4 items | 2 testlets of 1 to 5 items | 69 | 26 minutes allowed on items and 1 minute for instruction |

Candidates were given 120 minutes to answer a total of 228 items from the five subtests. There were seven groups of candidates who took a SEN version of the exam, and thus had extra time allowances in 2024. The timing and scoring of the SEN exams are explored in detail in Section 4.2.

In recent years, the mean scaled scores for QR and AR have been significantly higher compared to the other subtests, whereas the mean scaled score for VR has been notably lower. To address this disparity, UCAT decided in 2023 to scale down QR and AR by 10 points each and scale up VR by 20 points, aiming to reduce the gap between the cognitive subtests while maintaining consistent total cognitive subtest scores. This approach proved effective and was repeated in 2024, with QR and AR again scaled down by 10 points each and VR again scaled up by 20 points.

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 804. Universities received the cognitive subtest scaled scores plus a total score: a simple sum of the four cognitive subtest scores ranging from 1,200 to 3,600. SJT scaled scores are further categorised into four bands. The bands are determined by scaled score ranges as defined in Table 2.

Table 2. SJT Band Scaled Score Range and Description

| Band | Scaled Score Range | Intended Band Proportions | Narrative |
|---|---|---|---|
| Band 1 | 662–900 | 22% | Those in Band 1 demonstrated an excellent level of performance, showing similar judgement in most cases to the panel of experts. |
| Band 2 | 604–661 | 38% | Those in Band 2 demonstrated a good, solid level of performance, showing appropriate judgement frequently, with many responses matching model answers. |
| Band 3 | 508–603 | 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 | 300–507 | 10% | The performance of those in Band 4 was low, with judgement tending to differ substantially from ideal responses in many cases. |

The 2024 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 34 candidates took the online version of the test (see Section 4.4).

# 4. Examination Results

## 4.1 Overall Exam Results

This report presents the examination results for the 37,913 candidates who sat the UCAT between 8 July 2024 and 26 September 2024. Candidate numbers increased annually from 2017 to 2021 but declined slightly between 2021 and 2023. This year saw a new record high in candidate volume, surpassing the previous peak in 2021. The changes in candidate volume over time are shown in Figure 1 below.

Figure 1. Candidate Volumes since 2017

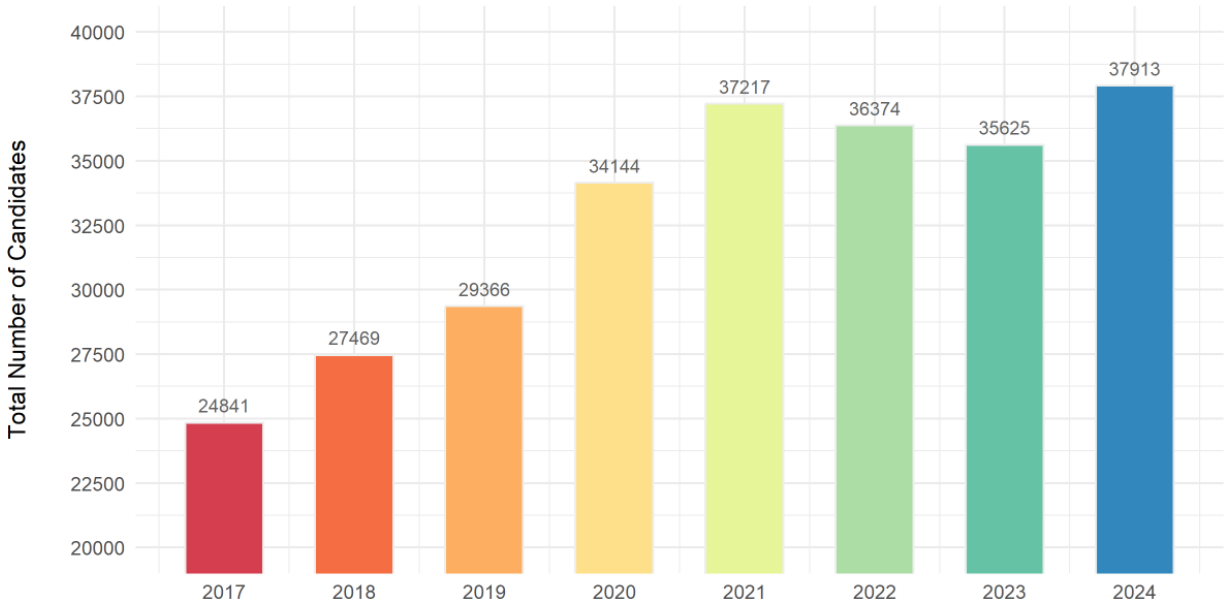


Table 3 presents summary statistics for each of the cognitive subtests plus the total scaled score for the cognitive subtests. VR scores were lowest with a mean score of 601, and the highest average score was achieved on AR with a mean of 653.

Table 3. Cognitive Subtest and Total Scaled Score Summary Statistics

| Subtest | Mean | *SD* | Min | Max |
|---|---|---|---|---|
| VR | 600.91 | 78.40 | 300 | 900 |
| DM | 619.99 | 91.01 | 300 | 900 |
| QR | 648.77 | 96.14 | 300 | 900 |
| AR | 653.48 | 102.88 | 300 | 900 |
| Total | 2,523.16 | 305.29 | 1,200 | 3,560 |

Figure 2 shows the change in scaled scores since 2017. The year 2017 was chosen as a starting point for comparison because prior to 2017 there was no operational DM section. Between 2018 and 2021, the mean scaled scores for VR, QR, and AR remained relatively stable. The significant drops in QR and DM scores in 2018 were attributed to changes in the scaling method for QR and the benchmark population for DM. Since 2017, AR scores have shown a continuous increase, possibly due to the relatively trainable nature of the AR subtest, resulting in a consistent upward trend as candidates and the public become more familiar with the subtest.

In 2022, a timing adjustment was introduced to alleviate the speededness of QR, which was expected to increase its average scaled score. To counterbalance this effect, QR was scaled down by 20 points. Despite this adjustment, QR and AR still had considerably higher average scaled scores than DM and VR, with VR being particularly lower than QR and AR in 2022. To address this imbalance, a decision was made to adjust the scaled scores in 2023 and 2024. QR and AR scores were reduced by 10 points each, while VR was increased by 20 points, with the aim of narrowing the gap between subtest scores while maintaining the total scaled score. This adjustment proved effective in 2023 and was effective to a lesser extent in 2024.

Despite being scaled down by 10 points, QR and AR displayed nearly identical average scaled scores to those in 2023. This suggests that upward-moving factors may have offset the downward scaling of QR and AR in 2024. For VR, although it was scaled up by 20 points, the average scaled score increased by only 10 points, indicating that downward-moving factors may have counteracted part of the upward scaling. These influences will be analysed further in the subsequent sections of this technical report, which will explore the factors contributing to the score changes observed this year.

Figure 2. Scaled Scores by Year since 2017

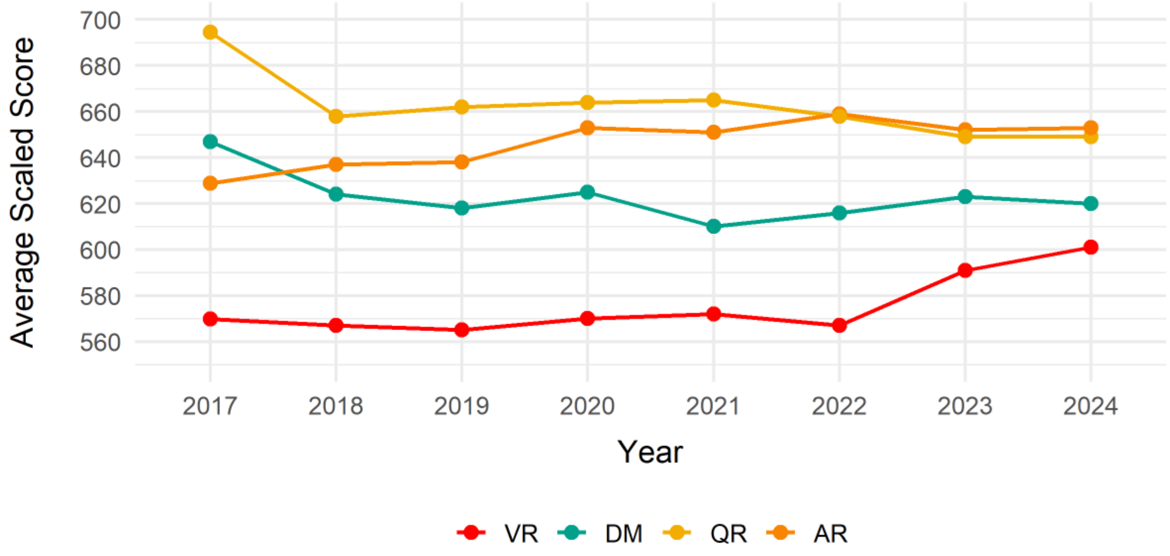


Table 4. Historic Cognitive Subtests Mean Scaled Scores (2017–2024)

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

Although the effects of the rescaling in 2024 were less pronounced than anticipated, the resulting score changes remain minor in the broader context. After accounting for the rescaling and considering the scenario if no rescaling was applied, the cohort-to-cohort deviations remain within 11 points for the subtests. These fluctuations are well below one *SEM* for these subtests, as outlined in Section 6. Statistically, such small deviations are not significant enough to raise concerns. This stability suggests consistent performance across cohorts, which aligns with expectations given the absence of major test changes and a relatively stable candidate composition. Overall, this indicates that the test produced this year functioned largely in line with those from previous years.

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

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

| | VR | DM | QR | AR |
|---|---|---|---|---|
| DM | 0.64*** | | | |
| QR | 0.58*** | 0.71*** | | |
| AR | 0.40*** | 0.56*** | 0.60*** | |
| SJT | 0.44*** | 0.50*** | 0.45*** | 0.43*** |

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

For the SJT, the number and percentage of candidates in each band for the 37,913 candidates who took the 2024 UCAT are shown in Table 6 below. Candidates are awarded a band for the SJT exam based on their underlying scaled score.

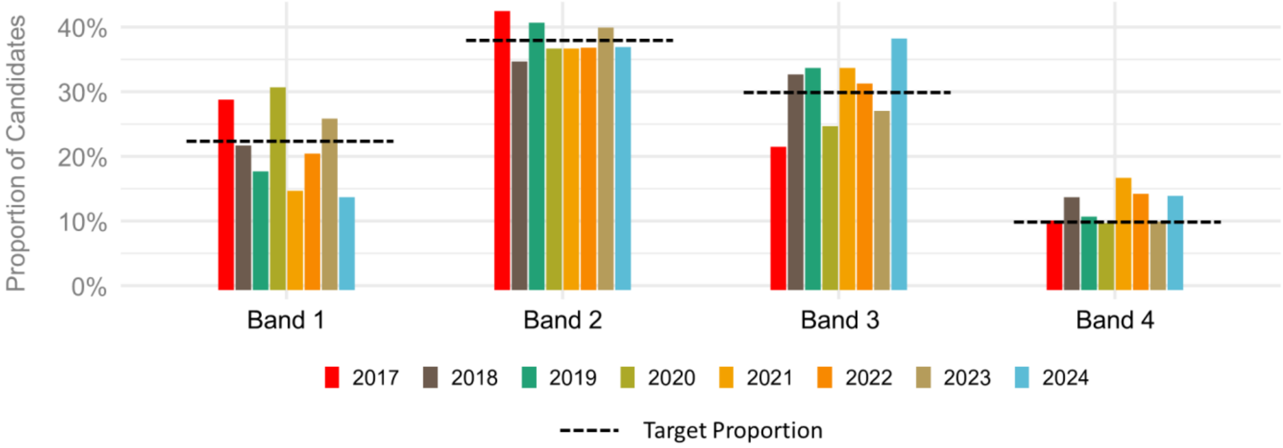
Table 6. SJT Band Distribution in 2024

| SJT Band | Number of Candidates | Mean Scaled Score | Percentage of Candidates | Target % |
|---|---|---|---|---|
| Band 1 | 4,933 | 680.68 | 13% | 22% |
| Band 2 | 13,736 | 631.67 | 36% | 38% |
| Band 3 | 14,237 | 564.84 | 38% | 30% |
| Band 4 | 5,007 | 444.24 | 13% | 10% |
| Total | 37,913 | 588.20 | 100% | 100% |

The proportions of candidates in the four bands deviated from the target distribution. This year, the deviations were particularly notable for Band 1 and Band 3, with deviations of 9% and 8%, respectively. Both higher bands, Band 1 and Band 2, had a lower-than-target proportion of candidates, while the lower bands, Band 3 and Band 4, had higher-than-expected proportions. This indicates that candidates in this cohort performed worse than anticipated, resulting in fewer candidates classified in Bands 1 and 2 and more candidates classified in Bands 3 and 4.

Figure 3 presents the distribution of candidates across SJT bands since 2017. Target proportions for each SJT band were introduced in 2018 and, while they vary slightly each year, they typically fluctuate within a 1% to 2% range. The 2024 target proportions are indicated by dotted lines in Figure 3. Although the deviation in SJT banding this year is larger than usual, it is not unprecedented when viewed in a historical context. While only 13% of candidates were classified in Band 1, it is close to the previous record low of 14% in 2021.

Figure 3. SJT Band Proportions 2017–2024



The deviation in SJT band distribution appears to follow a cyclical pattern. Cohorts with a higher-than-target proportion of candidates in high-performing bands are often followed by cohorts with fewer candidates than the target in those bands. While this pattern is not absolute, it is a noticeable trend. This cyclical behaviour aligns with the annual adjustment of banding cut-offs based on the performance of the previous cohort. A high-performing cohort raises the cut-off thresholds for the following year, often resulting in fewer candidates being classified in Bands 1 and 2 and more candidates being classified in Bands 3 and 4. While natural year-to-year performance fluctuations can occasionally offset this trend, they can also amplify it. For instance, a very strong cohort followed by another strong cohort may obscure the pattern, whereas a weaker cohort following a strong one can exacerbate it. The latter could be the case this year.

In 2023, a higher-than-target proportion of candidates was classified in both Band 1 and Band 2. This cyclical pattern partially explains the lower-than-target proportions observed in these bands in 2024. However, other factors, such as changes in candidate composition, may have also played a role.

A simulation using historic data was conducted to further investigate the cyclical pattern of SJT band distribution. The simulation showed that fixed banding cutoffs would substantially reduce the cyclical fluctuation across years. Using fixed cutoffs to stabilise band distribution is being considered for implementation in the future.

## 4.2 Special Educational Needs

There are seven exam versions available for SEN candidates who are granted extra time and breaks. However, only two candidates took the UCATSEN100 exam version, and 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 | 00:21:00 | 00:26:15 | 00:26:15 | 00:31:30 |
| DM | 00:31:00 | 00:38:45 | 00:38:45 | 00:46:30 |
| QR | 00:25:00 | 00:31:15 | 00:31:15 | 00:37:30 |
| AR | 00:12:00 | 00:15:00 | 00:15:00 | 00:18:00 |
| 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 | 00:31:30 | 00:42:00 | 00:42:00 | 00:21:00 |
| DM | 00:46:30 | 01:02:00 | 01:02:00 | 00:31:00 |
| QR | 00:37:30 | 00:50:00 | 00:50:00 | 00:25:00 |
| AR | 00:18:00 | 00:24:00 | 00:24:00 | 00:12:00 |
| SJT | 00:39:00 | 00:52:00 | 00:52:00 | 00:26:00 |

Only 6% of candidates took a SEN version of the exam, which is consistent with 2023. 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 | 35,565 | 94% |
| UCATSEN | 1,454 | 4% |
| UCATSENSA | 590 | 2% |
| UCATSEN50 | 64 | 0% |
| UCATSEN50SA | 41 | 0% |
| UCATSEN100 | 2 | 0% |
| UCATSEN100SA | 1 | 0% |
| UCATSA | 196 | 1% |
| Total | 37,913 | 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 four subtests. The sample sizes of UCATSEN50, UCATSEN50SA, and UCATSA are small and results for those versions should be treated with caution.

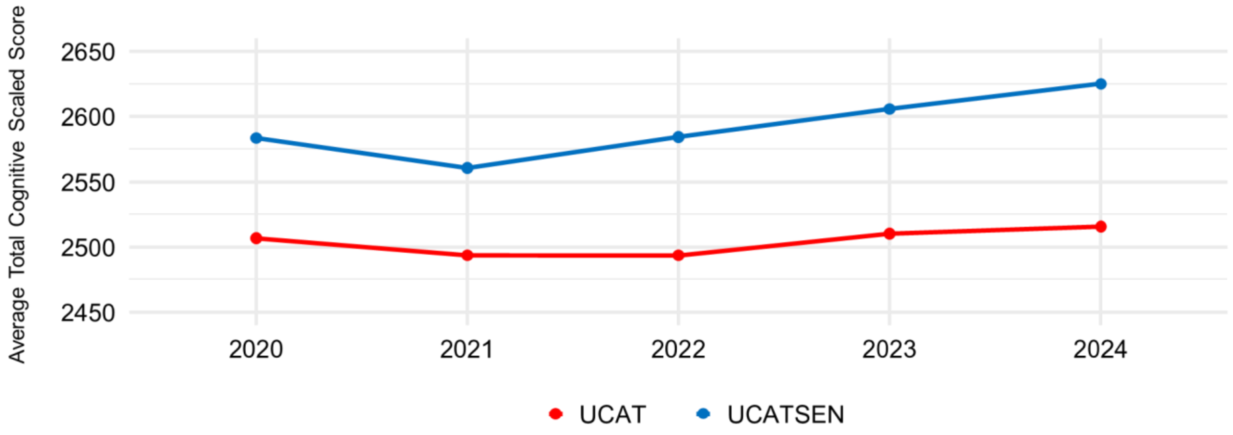
Table 10. SEN and Non-SEN Cognitive Subtests

| Subtest | Statistic | UCAT (35,565) | UCATSEN (1,454) | UCATSEN SA (590) | UCATSEN 50 (64) | UCATSEN 50SA (41) | UCATSA (196) |
|---|---|---|---|---|---|---|---|
| VR | Mean | 599.03 | 622.46 | 648.17 | 646.72 | 620.00 | 621.33 |
| | *SD* | 77.94 | 78.19 | 83.43 | 86.85 | 75.53 | 69.45 |
| | Min | 300 | 360 | 300 | 470 | 470 | 470 |
| | Max | 900 | 900 | 900 | 880 | 780 | 860 |
| DM | Mean | 618.14 | 646.55 | 658.00 | 630.94 | 630.98 | 638.78 |
| | *SD* | 91.07 | 85.88 | 84.71 | 76.48 | 95.13 | 82.92 |
| | Min | 300 | 300 | 370 | 420 | 440 | 440 |
| | Max | 900 | 890 | 900 | 780 | 870 | 840 |
| QR | Mean | 647.46 | 666.48 | 677.34 | 675.94 | 650.24 | 659.59 |
| | *SD* | 96.43 | 89.13 | 90.26 | 90.48 | 95.54 | 84.74 |
| | Min | 300 | 420 | 460 | 520 | 500 | 490 |
| | Max | 900 | 900 | 900 | 900 | 890 | 900 |
| AR | Mean | 651.15 | 689.67 | 690.10 | 691.09 | 691.22 | 676.94 |
| | *SD* | 102.28 | 104.55 | 108.67 | 111.93 | 118.41 | 96.49 |
| | Min | 300 | 450 | 300 | 490 | 480 | 440 |
| | Max | 900 | 900 | 900 | 900 | 900 | 900 |
| Total | Mean | 2515.77 | 2625.16 | 2673.61 | 2644.69 | 2592.44 | 2596.63 |
| | *SD* | 305.16 | 282.55 | 290.86 | 287.65 | 318.16 | 266.47 |
| | Min | 1,200 | 1,760 | 1,670 | 2,050 | 1,960 | 2,060 |
| | Max | 3,560 | 3,550 | 3,490 | 3,360 | 3,240 | 3,280 |

*Note.* UCATSEN100 and 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 presents the mean total cognitive scaled score for each exam version. It is evident that SEN candidates performed better than non-SEN candidates on the cognitive subtests overall. The difference in mean total cognitive scaled scores between candidates who sat the UCAT and those who sat the UCATSEN is 109 points. This is higher than the difference recorded in 2023 (95 points) and in 2022 (91 points). Figure 4 illustrates the differences in average total cognitive scaled scores between the UCAT and UCATSEN exam versions. While the gap between the two exam codes has widened slightly, the differences remain relatively consistent when viewed in the context of the total cognitive scaled scores, which usually range from 2,400 to 2,700 for these exam versions.

Figure 4. Average Total Cognitive Scaled Score: UCAT vs UCATSEN



The pattern of SEN candidates outperforming non-SEN candidates is also evident in the SJT results. The UCAT 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 UCATSEN100 and UCATSEN100SA versions are not disclosed for privacy reasons, as only two candidates and one candidate, respectively, sat these versions. Among the remaining exam versions, candidates performed best on the UCATSEN50SA, where 73% 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 | 586.36 | 13% | 36% | 38% | 14% |
| UCATSEN | 613.15 | 19% | 43% | 33% | 5% |
| UCATSENSA | 621.26 | 24% | 43% | 30% | 3% |
| UCATSEN50 | 614.22 | 19% | 38% | 39% | 5% |
| UCATSEN50SA | 625.54 | 34% | 39% | 17% | 10% |
| UCATSA | 619.08 | 18% | 48% | 31% | 3% |

*Note.* UCATSEN100 and 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, 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.

### 4.2.1 Pause-the-clock Accommodation

Since 2023, updated arrangements have been introduced for SEN candidates. Previously, such candidates were provided with an additional four minutes of rest time before the start of each section. Combined with the one minute available to all candidates during the introduction screen, this gave them a total of five minutes per section introduction. Starting in 2023, a “pause-the-clock” feature was implemented to provide greater flexibility, allowing candidates to use their additional rest time at any point during the test without restrictions. This change has proven effective, as candidates paused the test at various points, indicating diverse needs and preferences for rest periods.

As in 2023, the majority of candidates utilised the pause-the-clock feature. However, a small proportion, approximately 12–20%, did not use the feature, as shown in Table 13. This is slightly higher than the 12–13% observed last year. This may suggest potential difficulties in using the feature or that it does not effectively support certain SEN candidates, leading them to see no benefit in using it.

Table 12. Pause-the-clock Usage Count

| | UCATSA | | UCATSENSA | | UCATSEN50SA | |
|---|---|---|---|---|---|---|
| | *N* | % | *N* | % | *N* | % |
| Clock Not Paused | 39 | 20% | 96 | 16% | 5 | 12% |
| Clock Paused Within Allowed Time | 155 | 79% | 484 | 82% | 35 | 85% |
| Clock Paused Beyond Allowed Time | 2 | 1% | 10 | 2% | 1 | 2% |

The pause time allocated to candidates varied across different exam series codes. UCATSA, UCATSENSA, UCATSEN50SA, and UCATSEN100SA candidates were given 20, 25, 30, and 40 minutes respectively. Interestingly, on average, candidates utilised only a small portion of their allocated pause time. UCATSA candidates paused for approximately 7 minutes on average, UCATSENSA candidates paused for around 9 minutes, and UCATSEN50SA candidates paused for about 15 minutes. A summary of the total pause time used by candidates for each exam series code is provided in Table 14.

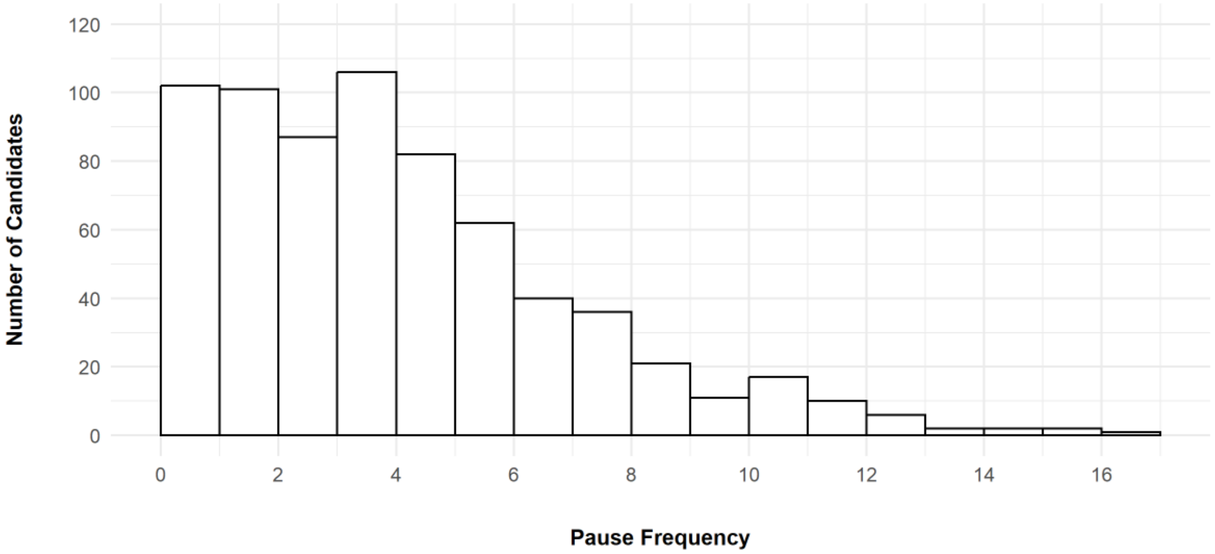
Table 13. Candidate Total Pause Time by Exam Series Code

| Exam Series Code | *N* | Candidate Total Pause Time (seconds) | | | | Total Allowed Pause Time (seconds) |
|---|---|---|---|---|---|---|
| | | Mean | *SD* | Min | Max | |
| UCATSA | 157 | 439.04 | 327.32 | 2.47 | 1,200 | 1,200 |
| UCATSENSA | 494 | 563.56 | 423.15 | 6.31 | 1,500 | 1,500 |
| UCATSEN50SA | 36 | 909.10 | 534.63 | 60.58 | 1,800 | 1,800 |

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

In 2023, a very small number of candidates were identified as potentially abusing the system by using the rest time to answer questions, with some pausing the test up to 105 times. To prevent such misuse, measures were introduced to ensure the pause feature was used solely for resting purposes. In 2024, candidates were limited to a maximum of three pauses per introduction section and three pauses per question section. This restriction significantly reduced the number of pauses, as illustrated in Figure 5. Most candidates paused between 1 and 6 times, with the highest recorded number of pauses being 17.

Figure 5. Candidate Pause Frequency Distribution



The frequency of pauses among candidates is outlined in Table 14. Up to 90% of candidates paused 8 times or fewer, with no significant evidence of potential misuse of the accommodation feature. The vast majority of candidates paused a reasonable number of times, with only 6% pausing more than 10 times. The most common number of pauses was 4, accounting for 15.4% of candidates, though this was not markedly higher than other frequencies. Approximately 15% of candidates paused once and approximately 15% paused twice. Those who paused between 1 and 5 times each represented over 10% of candidates, collectively comprising nearly 70% of the total. This relatively even distribution reflects the diverse needs for pause frequency among candidates. Some preferred more frequent pauses, while others opted for fewer, demonstrating the flexibility and effectiveness of the pause-the-clock feature compared to the previously fixed extended rest time in the introduction screen.

Table 14. Pause-the-clock Count Distribution

| Number of Pauses | *N* | % of Candidates | Cumulative % |
|---|---|---|---|
| 1 | 102 | 14.83% | 15% |
| 2 | 101 | 14.68% | 30% |
| 3 | 87 | 12.65% | 42% |
| 4 | 106 | 15.41% | 58% |
| 5 | 82 | 11.92% | 69% |
| 6 | 62 | 9.01% | 79% |
| 7 | 40 | 5.81% | 84% |
| 8 | 36 | 5.23% | 90% |
| 9 | 21 | 3.05% | 93% |
| 10 | 11 | 1.6% | 94% |
| 11 | 17 | 2.47% | 97% |
| 12 | 10 | 1.45% | 98% |
| 13 | 6 | 0.87% | 99% |
| 14 | 2 | 0.29% | 99% |
| 15 | 2 | 0.29% | 100% |
| 16 | 2 | 0.29% | 100% |
| 17 | 1 | 0.15% | 100% |

It is surprising to note that candidates who paused fewer times did not pause for longer durations per pause. Candidates who paused once, twice, or three times did not take longer pauses than those who paused more frequently. Regardless of the total number of pauses taken throughout the test, the average duration of each pause was approximately two minutes. Therefore, candidates who paused more often accumulated a longer total pause time. This observation could provide valuable insights for future accommodation arrangements. The pause time by the number of pauses is summarised in Table 15.

Table 15. Average Pause Time by Number of Pauses

| Number of Pauses | *N* | Average Item Pause Time (sec) | Average Total Pause Time (sec) |
|---|---|---|---|
| 1 | 102 | 142.12 | 142.12 |
| 2 | 101 | 122.44 | 244.88 |
| 3 | 87 | 122.00 | 366.01 |
| 4 | 106 | 159.37 | 637.49 |
| 5 | 82 | 138.32 | 691.59 |
| 6 | 62 | 119.39 | 716.33 |
| 7 | 40 | 127.91 | 895.38 |
| 8 | 36 | 112.04 | 896.29 |
| 9 | 21 | 91.20 | 820.83 |
| 10 | 11 | 102.95 | 1,029.46 |
| 11 | 17 | 107.50 | 1,182.45 |
| 12 | 10 | 80.80 | 969.54 |
| 13 | 6 | 103.80 | 1,349.34 |
| 14 | 2 | 78.86 | 1,104.01 |
| 15 | 2 | 97.25 | 1,458.70 |
| 16 | 2 | 64.36 | 1,029.70 |
| 17 | 1 | 56.97 | 968.47 |

A detailed breakdown of pause usage is provided in Table 17. On average, candidates who utilised the pause feature paused 4.55 times throughout the test, with an average total pause duration of approximately nine minutes (555.9 seconds). Similar to last year, the majority of pauses occurred during the middle of the exam, particularly in the DM and QR subtests. However, the variation in the number of pauses across subtest sections was less pronounced this year. In 2024, the range of pauses per section was more evenly distributed, from 223 to 581 pauses, compared to the wider range of 125 to 794 observed in 2023. This narrower range and more evenly distributed rests may show more genuine rest patterns, as the pause restrictions likely eliminated strategic misuse of pause time for answering questions.

On average, pauses during the introduction screens were longer than those within subtests. However, more pauses were recorded during the subtests than the introduction screens. As observed last year, almost no candidates paused during the VR introduction screen, with only six pauses recorded. This is expected, as the VR section is the first part of the test, and candidates are less likely to require rest at that point. This highlights the benefit of the pause-the-clock feature, as it allows candidates to redistribute unused pause time to later sections when it is more needed. This flexibility clearly benefits candidates and demonstrates the effectiveness of the feature.

Table 16. Pause-the-clock Overall Usage in Each Section

| Section | Total Pause Count (across all candidates) | Pause Usage per Individual Candidate | | | | |
|---|---|---|---|---|---|---|
| | | Pause Count | | Time Paused (seconds) | | |
| | | Mean | Max | Mean | Minimum | Maximum |
| UCAT Exam Introduction | 4 | 1 | 1 | 103.59 | 10.05 | 241.47 |
| Pause the Clock Introduction | 1 | 1 | 1 | 0.97 | 0.97 | 0.97 |
| Verbal Reasoning Introduction | 6 | 1.20 | 2 | 64.07 | 2.50 | 106.67 |
| Verbal Reasoning Subtest | 392 | 1.50 | 3 | 66.94 | 1.19 | 321.56 |
| Decision Making Introduction | 223 | 1.01 | 2 | 148.65 | 1.61 | 711.70 |
| Decision Making Subtest | 581 | 1.58 | 3 | 94.72 | 0.91 | 568.39 |
| Quantitative Reasoning Introduction | 287 | 1.03 | 3 | 192.03 | 6.70 | 826.42 |
| Quantitative Reasoning Subtest | 409 | 1.44 | 3 | 108.17 | 0.88 | 748.77 |
| Abstract Reasoning Introduction | 295 | 1.05 | 3 | 183.75 | 2.45 | 919.95 |
| Abstract Reasoning Subtest | 247 | 1.31 | 3 | 95.34 | 1.66 | 702.77 |
| Situational Judgement Introduction | 216 | 1.04 | 2 | 149.15 | 1.19 | 760.67 |
| Situational Judgement Subtest | 359 | 1.50 | 3 | 117.01 | 2.55 | 710.11 |
| Reviewing Screen | 110 | 1.47 | 4 | 144.56 | 6.58 | 553.89 |
| **Total** | **3,130** | **4.55** | **17** | **555.90** | **2.47** | **2,400** |

In summary, the trial of the pause-the-clock accommodation has demonstrated its advantages in offering candidates greater flexibility. The second-year implementation, with the addition of pause restrictions, effectively resolved issues of misuse. However, the feature remains slightly underutilised, with a proportion of candidates not using it despite being eligible. Further investigation is needed to determine whether this is due to difficulties in usage or because the accommodation does not align with the needs of these candidates. Additionally, most candidates did not use the full allocated pause time, suggesting that the amount of pause time provided could be reviewed and adjusted if necessary. Ongoing monitoring and analysis of pause patterns will be crucial to refining these measures and ensuring they continue to meet the needs of all candidates in a fair and equitable manner.

## 4.3 Medicine and Dentistry

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

The majority of candidates applied for medicine, accounting for 55% of the total number of candidates. While applicants for medicine remain the largest group and comprise more than half of all candidates, there has been a gradual decline in medicine candidates, down from 59% in 2023, 63% in 2022, and 69% in 2021. In contrast, 13% of candidates applied for dentistry, consistent with 2023, but an increase from 11% in 2022 and 9% in 2021. The remaining 32% either applied for other courses or could not be matched with UCAS data. This is a slight rise from 29% in 2023, 26% in 2022, and 23% in 2021, potentially due to an increasing number of partner international universities outside the UCAS system.

Candidates who applied for medicine as a first choice outperformed those who applied for dentistry, as illustrated in Table 17. The highest mean scaled score was achieved on AR and the lowest on VR for both candidate groups. Candidates who did not apply for medicine or dentistry or were not matched by UCAS data performed less well than both other groups.

Table 17. Medicine/Dentistry Candidates: Cognitive and Total Scaled Scores

| Subtest | Mean | | | *SD* | | |
|---|---|---|---|---|---|---|
| | Medicine | Dentistry | None | Medicine | Dentistry | None |
| VR | 619.60 | 603.36 | 567.66 | 76.45 | 67.60 | 74.82 |
| DM | 644.23 | 631.68 | 573.39 | 86.01 | 80.41 | 85.50 |
| QR | 672.27 | 664.25 | 601.89 | 93.99 | 87.50 | 85.70 |
| AR | 678.32 | 676.28 | 601.33 | 101.38 | 98.74 | 86.34 |
| Total | 2,614.42 | 2,575.57 | 2,344.27 | 287.51 | 263.93 | 271.08 |

Better performance by medicine candidates is also evident in the SJT banding. As shown in Table 18, 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 and Band 2 than dentistry candidates, though the difference is minimal and the split is comparable.

Table 18. Medicine/Dentistry Candidates: SJT Bands

| Group | Mean Scaled Score | Band 1 | Band 2 | Band 3 | Band 4 |
|---|---|---|---|---|---|
| Dentistry | 605.64 | 16% | 42% | 35% | 6% |
| Medicine | 607.31 | 17% | 43% | 35% | 6% |
| None | 548.10 | 6% | 23% | 43% | 29% |

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

## 4.4 Mode of Delivery

In 2024, the UCAT was offered in both the standard test centre and online proctored mode. Only 34 candidates took the exam in the online proctored mode, amounting to only 0.09% of all candidates. This contrasts with 2020, when more than 11,038 candidates took the exam in the online mode. The proportion of candidates using the online version of the test is decreasing as test centres are back open fully and candidates are encouraged to use a test centre where possible.

Given the large difference in volumes between the two modes and the low number of candidates who took the test in the online mode in 2024, it is not possible to draw reliable inferences on differences in performance for the 2024 cohort of candidates.

## 4.5 Examination Results by Demographic Variables

### 4.5.1 Variation by Demographic Group

Pearson VUE undertakes several tasks as part of the item development and analysis process to ensure differential performance related to demographic characteristics are not caused by the test content or mode of delivery. All content creators and reviewers complete an editorial course and agree to a global set of principles and best practices that need to be considered when creating content. Item writers and editors are provided with specific guidelines to be adhered to when creating content. Test items are developed using a group of content creation specialists, and bias, sensitivity, and accessibility reviews are undertaken before test items are used in the exam. We also produce practice resources that are freely accessible to all. Finally, we analyse the performance of individual items by demographic characteristic and remove any items that might exhibit bias (as discussed in Section 7.3).

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

### 4.5.2 Gender

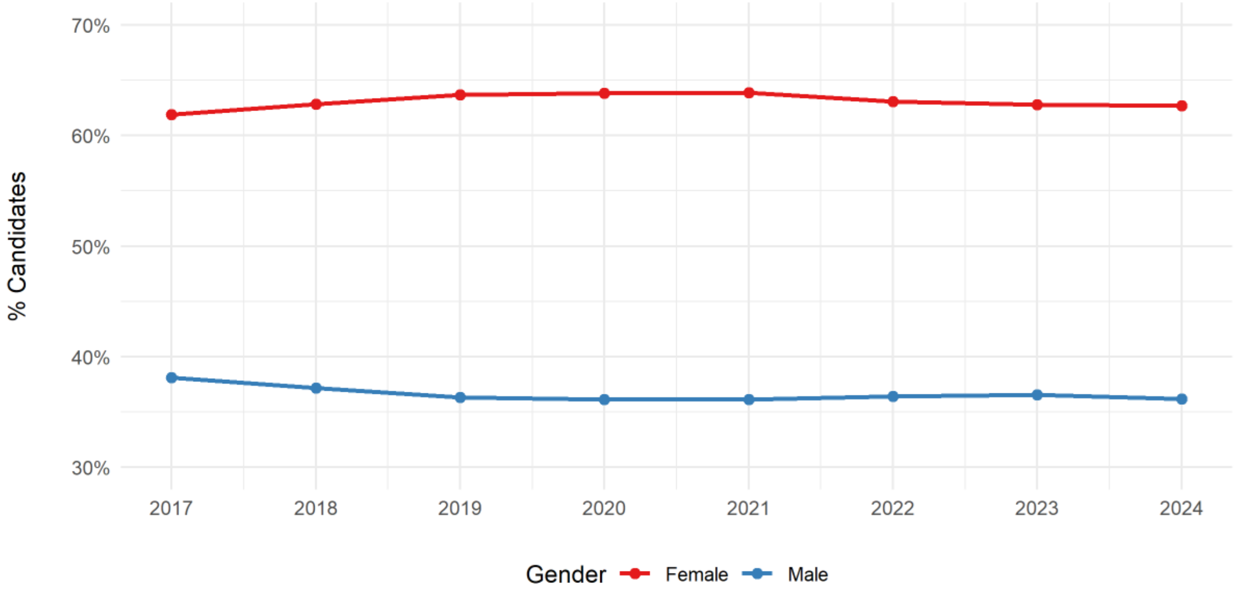
Table 19 provides the breakdown of test-takers by preferred gender term. The majority of test-takers identified as female, while only 433 indicated that they "use another term" to describe their gender or preferred not to disclose their gender.

Table 19. Gender Counts

| Gender | *N* | % |
|---|---|---|
| Female | 23,765 | 63% |
| Male | 13,715 | 36% |
| I prefer not to say | 365 | 1% |
| I use another term | 68 | 0% |

The distribution of candidates by gender has remained stable since 2017, with a slight increase in female candidates from 2017 to 2019 (Figure 6).

Figure 6. Distribution of Candidates by Gender 2017–2024



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. presents the differences in average scores between male and female candidates.

Table 20 presents the differences in average scores between male and female candidates.

Table 20. Gender Scaled Scores

| Subtest | Mean Scaled Score | | *SD* Scaled Score | |
|---|---|---|---|---|
| | Female | Male | Female | Male |
| VR | 595.50 | 609.17 | 77.15 | 79.01 |
| DM | 611.57 | 633.87 | 89.49 | 91.71 |
| QR | 635.02 | 672.25 | 91.49 | 99.32 |
| AR | 647.23 | 664.06 | 100.35 | 106.04 |
| Total Cognitive | 2,489.32 | 2,579.36 | 296.88 | 310.28 |
| SJT | 593.94 | 577.93 | 72.77 | 76.19 |

A statistical test was used to examine whether the differences between the two groups observed in Table 20 were statistically significant. Table 21 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.

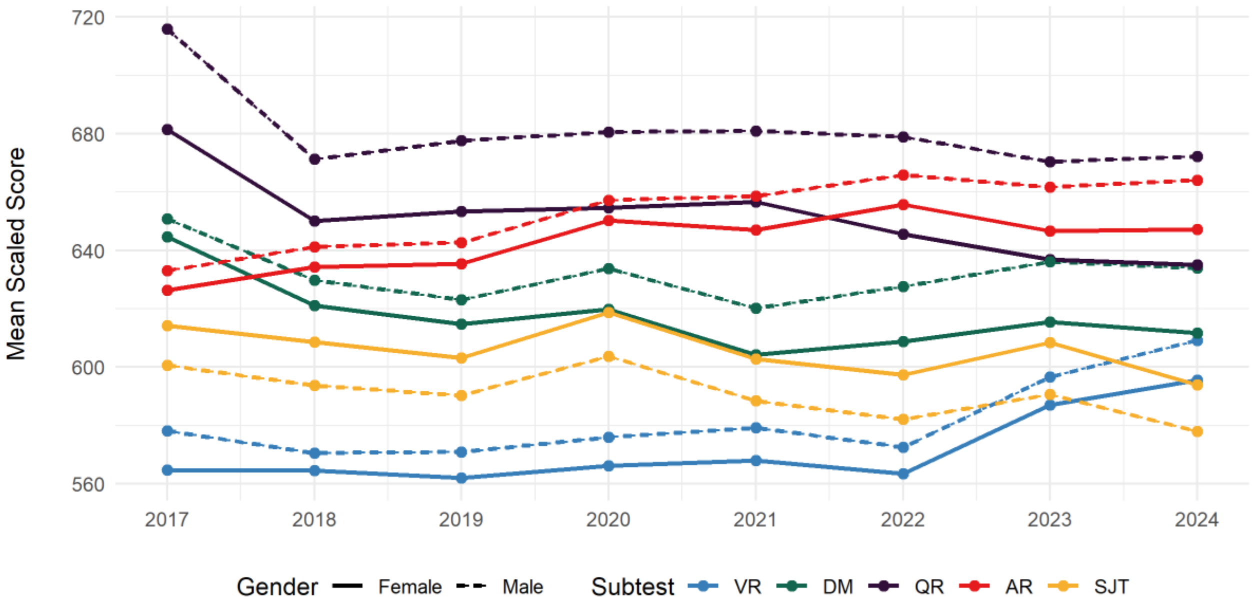
Therefore, Table 21 shows us that there are differences between male and female performance on each subtest and on the total cognitive scores, and that these differences are likely not to be the result of random chance.

Table 21. Gender *t*-Test

| Subtest | *t*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 16.38 | 37,478 | < 0.01 |
| DM | 23.03 | 37,478 | < 0.01 |
| QR | 36.77 | 37,478 | < 0.01 |
| AR | 15.32 | 37,478 | < 0.01 |
| Total Cognitive | 27.82 | 37,478 | < 0.01 |
| SJT | -20.17 | 37,478 | < 0.01 |

Figure 7 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 and AR subtests.

Figure 7. Scaled Score Distribution of Candidates by Gender 2017–2024



## 4.5.3 Ethnicity

UCAT candidates who reside in the UK are requested to answer a question relating to their ethnicity. The ethnic categories in the questionnaire were simplified in 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

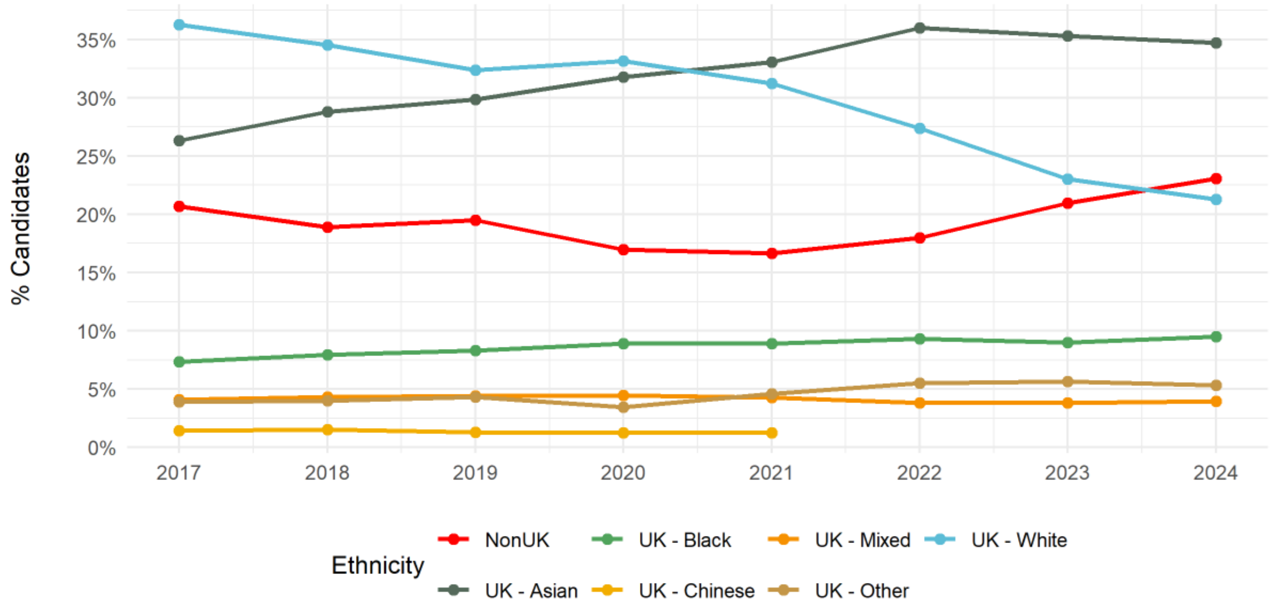
Table 22 shows the breakdown of candidates by ethnicity in the 2024 exam. The biggest candidate group was UK-Asian. Twenty-four percent of candidates were not categorised due to being non-UK candidates.

Table 22. Ethnic Group Counts

| Country | Ethnic Group | *N* | % UK Candidates | % Total Candidates |
|---|---|---|---|---|
| UK | Asian | 13,165 | 46% | 35% |
| UK | White | 8,073 | 28% | 22% |
| UK | Black | 3,608 | 13% | 10% |
| UK | Other ethnic group | 2,019 | 7% | 5% |
| UK | Mixed | 1,493 | 5% | 4% |
| Non-UK | Non-UK | 8,751 | - | 24% |

The proportion of candidates across most ethnic groups has remained relatively stable in recent years, with a few notable exceptions. Figure 8 highlights the trends in the ethnic composition of the candidate pool. Since 2017, the proportion of UK-Asian candidates has gradually increased, while the proportion of UK-White candidates has steadily declined. Since 2021, UK-Asian has become the largest ethnic group in the sample, overtaking UK-White. There has also been a gradual increase in non-UK candidates since 2022. This year, the UK-White group dropped to the third largest ethnic group in the sample, with the non-UK group rising to the second largest.

Figure 8. Distribution of Candidates by Ethnic Group 2017–2024



UK-White candidates achieved the highest average scores across all subtests compared to other ethnic groups. Table 23 provides a breakdown of the average scores for each subtest by ethnic group. UK-Black candidates had the lowest average performance in DM, QR, AR, 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 23. Ethnic Group Mean Scaled Score

| Subtest | White | Asian | Black | Mixed | Other Ethnic Group | Non-UK |
|---|---|---|---|---|---|---|
| VR | 623.01 | 600.31 | 584.40 | 617.64 | 577.59 | 590.17 |
| DM | 647.44 | 619.19 | 587.57 | 637.25 | 601.50 | 611.16 |
| QR | 661.11 | 658.72 | 608.41 | 655.71 | 637.37 | 641.24 |
| AR | 669.15 | 664.97 | 616.88 | 667.03 | 647.43 | 636.35 |
| Total Cognitive | 2,600.72 | 2,543.19 | 2,397.27 | 2,577.62 | 2,463.90 | 2,478.92 |
| SJT | 608.27 | 596.75 | 586.17 | 602.84 | 583.42 | 555.31 |

An *F*-test was used to examine whether the differences observed in Table 23 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 24 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 23 are likely to reflect true differences in performance in the candidate population.

Table 24. Ethnic Group *F*-Test

| Subtest | *F*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 209.82 | 6 | < 0.01 |
| DM | 245.20 | 6 | < 0.01 |
| QR | 172.17 | 6 | < 0.01 |
| AR | 186.17 | 6 | < 0.01 |
| Total Cognitive | 260.25 | 6 | < 0.01 |
| SJT | 457.14 | 6 | < 0.01 |

Figure 9 presents the mean total cognitive scaled scores from 2017 to 2024. The ranking of ethnic groups has remained relatively consistent over time, with no significant changes apart from a minor intersection between the non-UK group and the Other Ethnic Group in 2022. The highest-performing group for the total cognitive scaled score has consistently been UK-White candidates, excluding the UK-Chinese group, which was removed from the survey in 2022. Following the UK-White candidates, the ranking is UK-Mixed, UK-Asian, non-UK, UK-Other, and UK-Black. This order has been very stable across the years and remains largely unchanged this year.

Figure 9. Ethnic Group Mean Scaled Score for Total Scaled Score 2017–2024

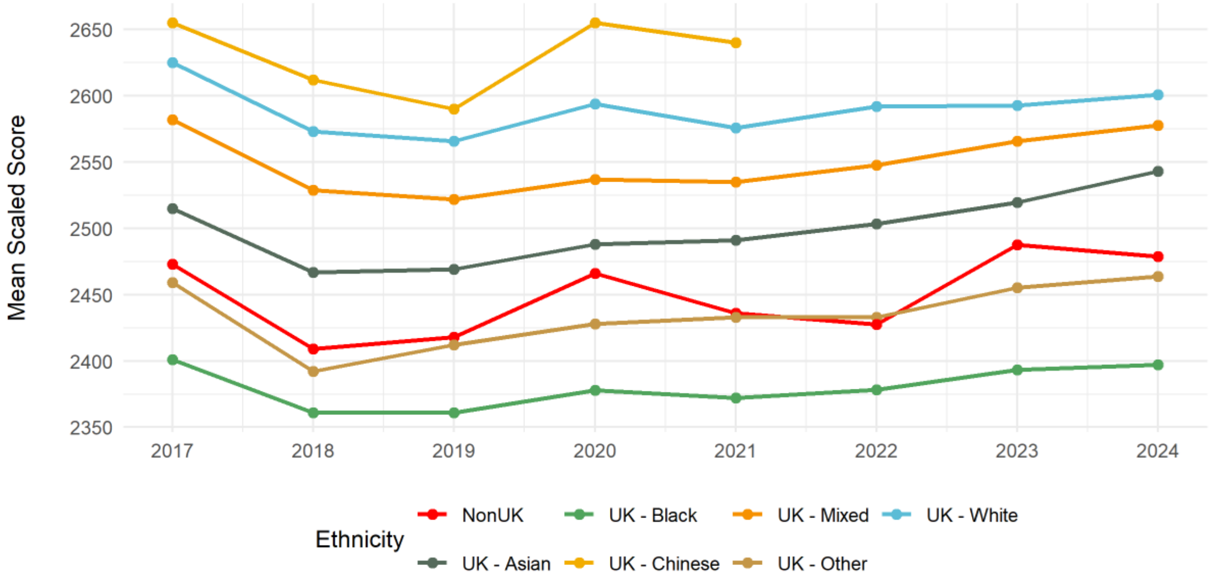
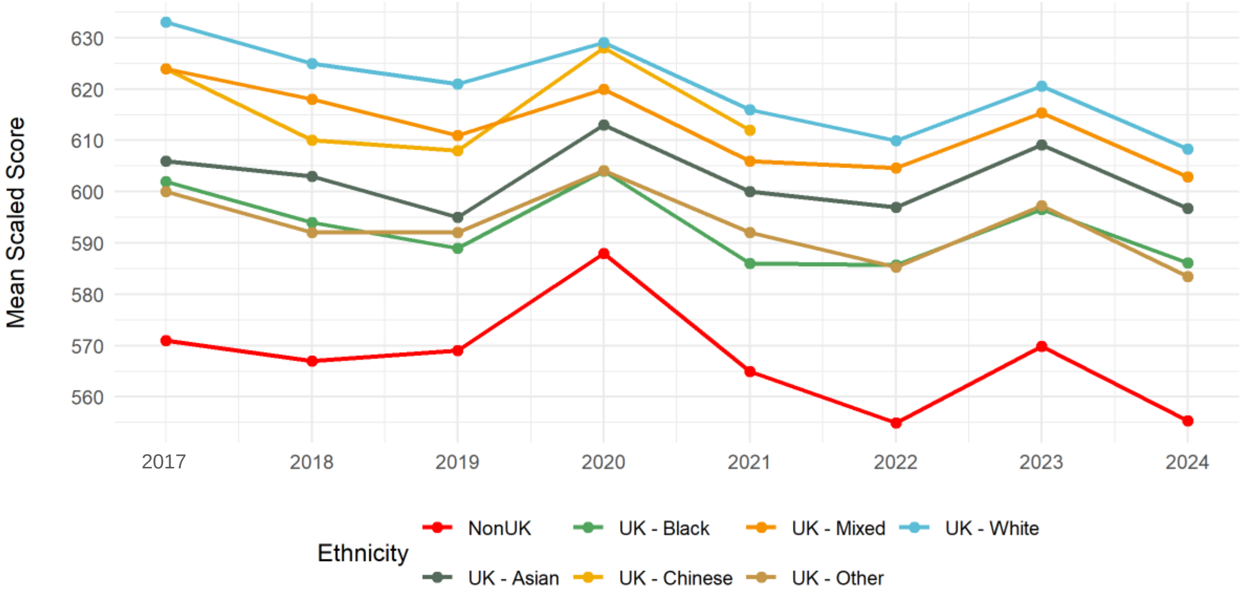


Figure 10 shows the mean scaled scores for the SJT by ethnic group from 2017 to 2024. While the ranking of ethnic groups for the SJT is relatively stable, it is less consistent compared to the total cognitive scaled scores, with several intersections occurring across the years. The order of performance for the SJT also differs slightly from that of the total cognitive scaled scores. For the SJT, non-UK candidates consistently perform the worst, with a clear margin separating them from other groups. This underperformance may be linked to a relationship between situational judgement and cultural competence, as UK-based candidates are more likely to have a better understanding of UK-specific situational norms and behaviours. However, it is important to emphasise that no evidence of bias against candidates based on residency has been identified at the item level within the SJT.

Figure 10. Ethnic Group Mean Scaled Score for SJT 2017–2024



## 4.5.4 Socio-Economic Classification (SEC)

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

This issue is illustrated in Table 25 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.

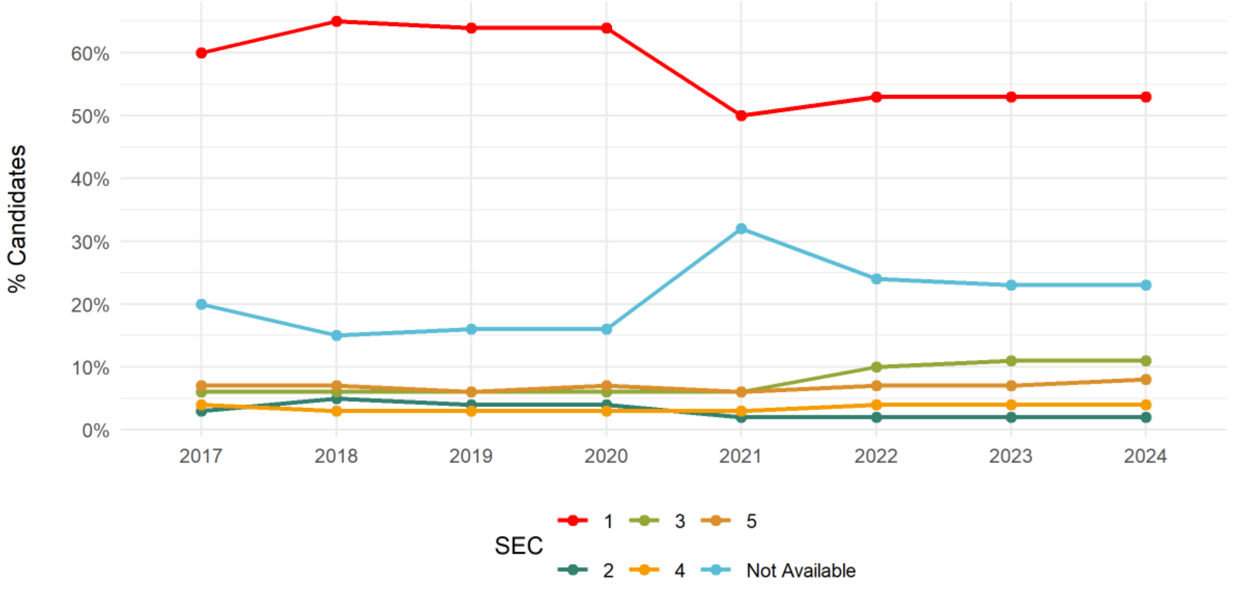
Table 25. SEC Counts

| Country | SEC | *N* | % of SEC | % of All |
|---|---|---|---|---|
| UK | 1 | 15,517 | 53% | 41% |
| | 2 | 541 | 2% | 1% |
| | 3 | 3,072 | 11% | 8% |
| | 4 | 1,069 | 4% | 3% |
| | 5 | 2,195 | 8% | 6% |
| | Unknown | 6,768 | 23% | 18% |
| EU | | 1,151 | | 3% |
| Other | | 7,600 | | 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 11. 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 2024 is similar to that observed in 2023.

Figure 11. Candidates by SEC 2017–2024



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 26.

Table 26. SEC Scaled Scores

| Mean Scaled Score | | | | | | |
|---|---|---|---|---|---|---|
| Subtest | SEC 1 | SEC 2 | SEC 3 | SEC 4 | SEC 5 | NA |
| VR | 615.07 | 606.82 | 599.45 | 596.28 | 586.38 | 587.98 |
| DM | 638.46 | 621.83 | 617.53 | 609.57 | 595.91 | 599.51 |
| QR | 665.60 | 643.18 | 647.06 | 636.75 | 627.58 | 629.90 |
| AR | 671.86 | 641.76 | 654.99 | 644.69 | 637.32 | 640.39 |
| Total Cognitive | 2,590.99 | 2,513.59 | 2,519.03 | 2,487.29 | 2,447.19 | 2,457.78 |
| SJT | 605.24 | 600.18 | 595.69 | 588.52 | 589.94 | 586.68 |
| *SD* | | | | | | |
| VR | 75.30 | 69.81 | 72.46 | 69.63 | 67.12 | 75.39 |
| DM | 86.92 | 85.00 | 84.20 | 83.36 | 80.87 | 89.11 |
| QR | 93.07 | 89.64 | 88.15 | 86.58 | 85.78 | 89.33 |
| AR | 101.83 | 94.35 | 98.03 | 93.96 | 95.01 | 99.02 |
| Total Cognitive | 289.93 | 271.91 | 276.12 | 271.82 | 265.03 | 291.69 |
| SJT | 61.59 | 68.27 | 66.13 | 67.80 | 68.04 | 72.96 |

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

Table 27. SEC *F*-Test

| Subtest | *F*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 161.93 | 5 | < 0.01 |
| DM | 249.28 | 5 | < 0.01 |
| QR | 189.40 | 5 | < 0.01 |
| AR | 127.68 | 5 | < 0.01 |
| Total Cognitive | 267.05 | 5 | < 0.01 |
| SJT | 89.69 | 5 | < 0.01 |

### 4.5.5 Age

The majority of UCAT candidates are aged 16–19 years old. A small minority of candidates are 35 or older and an even smaller proportion are under 16 (Table 28). A steady proportional increase in candidates aged 16–19 taking the test can be observed: 76% of the testing population was aged 16–19 in 2020, 78% in 2021, 81% in 2022, 82% in 2023, and it continued to be 82% in 2024.

Table 28. Age Counts

| Age | *N* | Percent |
|---|---|---|
| ≤ 15 | 59 | 0% |
| 16–19 | 30,947 | 82% |
| 20–24 | 5,066 | 13% |
| 25–34 | 1,486 | 4% |
| ≥ 35 | 336 | 1% |

Candidates who were aged 16–19 tended to perform better in all cognitive subtests, as illustrated in Figure 12 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 13.

Figure 12. Mean Scaled Scores by Age

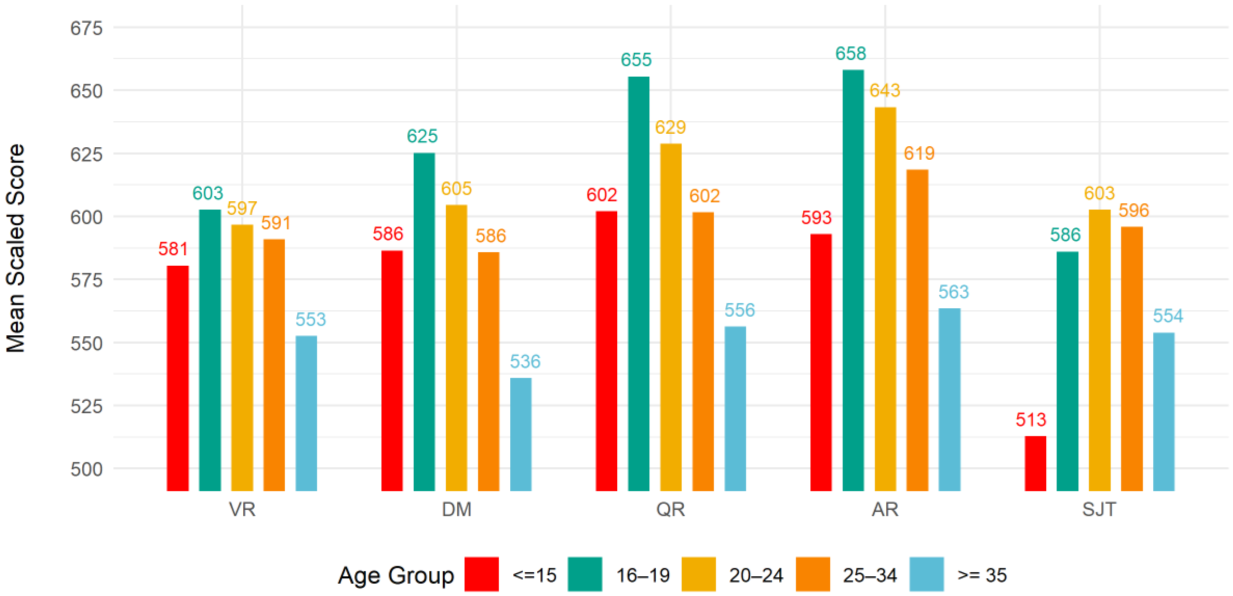
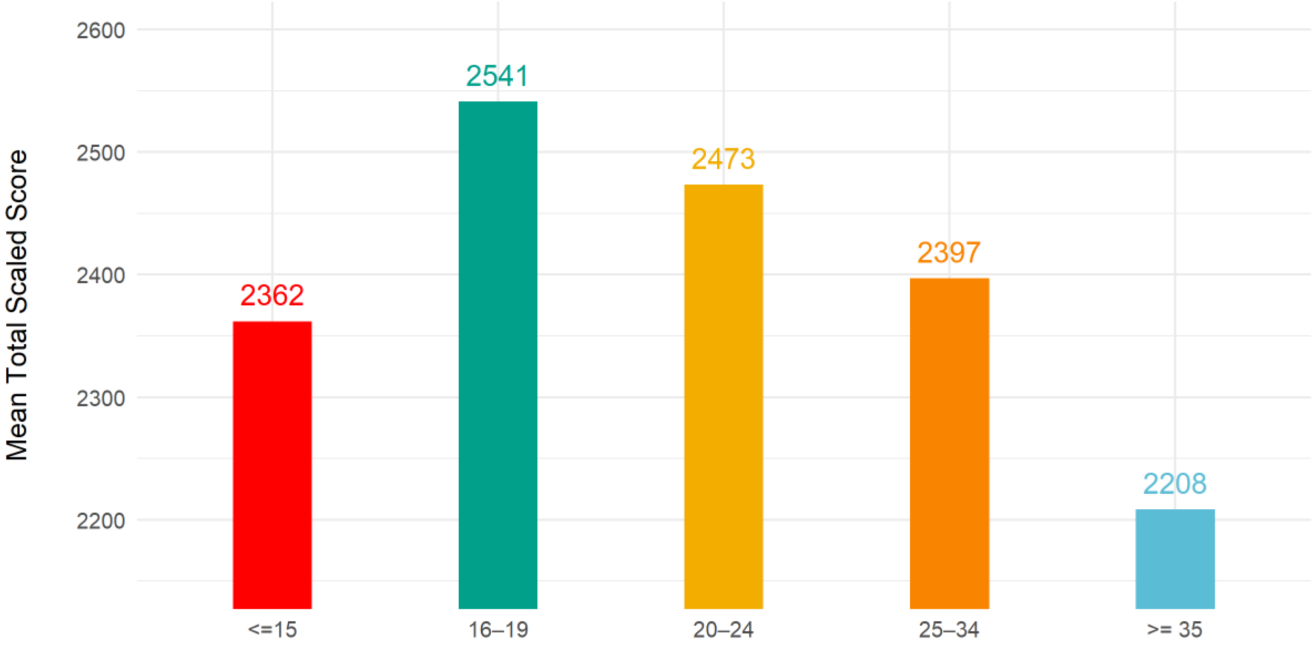


Figure 13. Mean Total Scaled Scores of Cognitive Subtests by Age



Hypothesis testing demonstrated that the differences observed among the groups is unlikely to have occurred due to chance, as shown in Table 29.

Table 29. Age *F*-Test

| Subtest | *F*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 46.66 | 4 | < 0.01 |
| DM | 192.14 | 4 | < 0.01 |
| QR | 270.05 | 4 | < 0.01 |
| AR | 141.79 | 4 | < 0.01 |
| Total | 223.47 | 4 | < 0.01 |
| SJT | 91.64 | 4 | < 0.01 |

To examine the relationship between age and subtest performance, Table 30 shows the correlations between candidate age and performance on each subtest. As highlighted in the significance column, all subtests exhibit statistically significant correlations. For the cognitive subtests, there is a slight negative correlation with age, indicating that younger candidates tended to perform better. This may reflect the possibility that most candidates take the test immediately after completing secondary school, whereas older candidates may include those who have taken alternative routes or experienced delays in entering medicine or dentistry, which could suggest they were initially less prepared or competitive. However, it is important to note that the correlation between age and cognitive subtest performance is small. In contrast, the correlation between age and performance on the SJT is positive. While this might appear to suggest that older candidates perform better on the SJT, the effect size of the correlation is extremely small and close to negligible.

Table 30. Correlation of Scaled Score with Age (ungrouped)

| Subtest | Correlation | Significance |
|---|---|---|
| VR | -0.08 | *p* < 0.01 |
| DM | -0.15 | *p* < 0.01 |
| QR | -0.17 | *p* < 0.01 |
| AR | -0.13 | *p* < 0.01 |
| Total Cognitive | -0.16 | *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.

### 4.5.6 Education

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

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

The majority of candidates in 2024 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 2023.

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 31. Table 32 shows that the differences observed in Table 31 are statistically significant.

Table 31. Education Scaled Scores

| Subtest | School Leaver Qualification | Degree Level or Above |
|---|---|---|
| Mean Scaled Score | | |
| *N* | 31,419 | 5,784 |
| VR | 603.04 | 593.30 |
| DM | 625.03 | 597.65 |
| QR | 655.26 | 618.98 |
| AR | 657.92 | 633.64 |
| Total Cognitive | 2,541.25 | 2,443.57 |
| SJT | 587.66 | 597.26 |
| *SD* | | |
| VR | 77.80 | 80.27 |
| DM | 90.34 | 89.56 |

| Subtest | School Leaver Qualification | Degree Level or Above |
|---|---|---|
| QR | 96.33 | 88.40 |
| AR | 102.61 | 101.16 |
| Total Cognitive | 303.10 | 298.36 |
| SJT | 72.61 | 77.52 |

Table 32. Education *t*-Test

| Subtest | *t*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | -8.71 | 37,201 | < 0.01 |
| DM | -21.21 | 37,201 | < 0.01 |
| QR | -26.65 | 37,201 | < 0.01 |
| AR | -16.57 | 37,201 | < 0.01 |
| Total Cognitive | -22.58 | 37,201 | < 0.01 |
| SJT | 9.14 | 37,201 | < 0.01 |

### 4.5.7 Country of Residence

Candidates were required to state their country of residence, and these are categorised as UK, EU or Rest of World. The majority of candidates who take the UCAT reside in the UK, as can be seen in Table 33 below.

Table 33. Candidate Count by Residence

| Country of Permanent Residence | *N* | Percent |
|---|---|---|
| UK | 29,162 | 77% |
| Rest of World | 7,600 | 20% |
| EU | 1,151 | 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 14.

Figure 14. Country of Residence 2017–2024

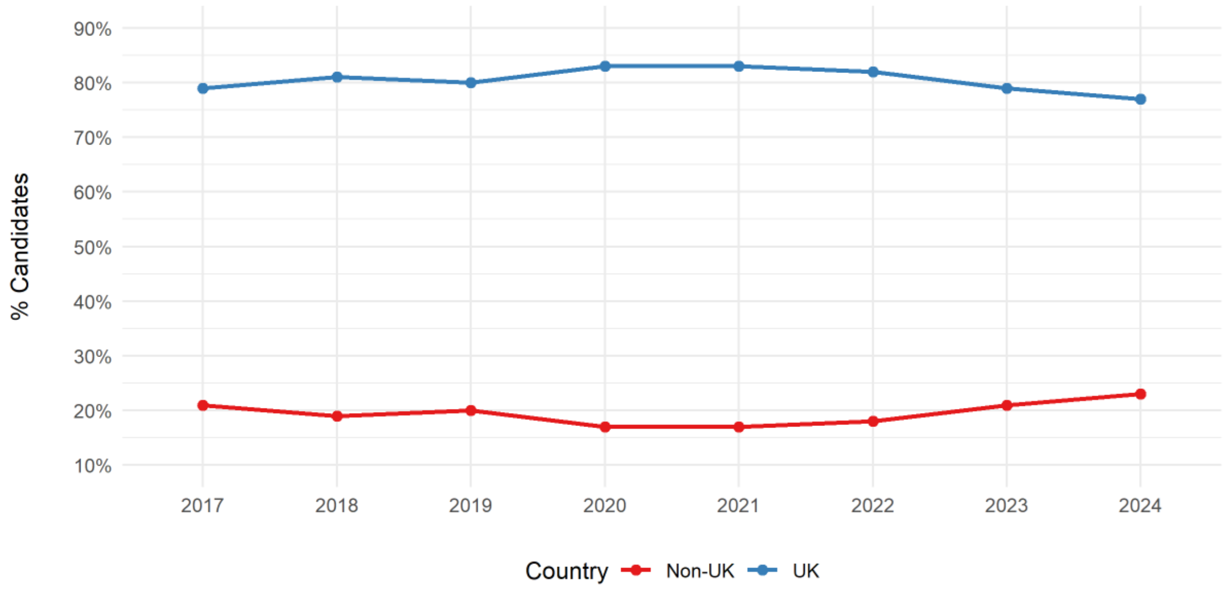


Table 34 indicates that UK candidates outperform EU and Rest of World candidates across all subtests. Rest of World candidates generally perform better than EU candidates in all subtests, except for the SJT.

Table 34. Candidate Scaled Scores by Residence

| Subtest | UK | Rest of World | EU |
|---|---|---|---|
| Mean Scaled Score | | | |
| VR | 604.14 | 590.60 | 587.34 |
| DM | 622.64 | 612.18 | 604.42 |
| QR | 651.02 | 645.46 | 613.41 |
| AR | 658.63 | 637.54 | 628.44 |
| Total Cognitive | 2,536.43 | 2,485.78 | 2,433.61 |
| SJT | 598.07 | 552.81 | 571.77 |
| *SD* | | | |
| VR | 75.15 | 89.16 | 75.78 |
| DM | 88.37 | 100.62 | 85.54 |
| QR | 92.33 | 110.16 | 83.02 |
| AR | 100.95 | 108.59 | 97.99 |
| Total Cognitive | 292.59 | 348.66 | 277.32 |
| SJT | 66.21 | 90.65 | 76.68 |

An *F*-test of the differences observed between UK and non-UK candidates is presented in Table 35 below. It shows that the differences are statistically significant.

Table 35. Residence *F*-Test

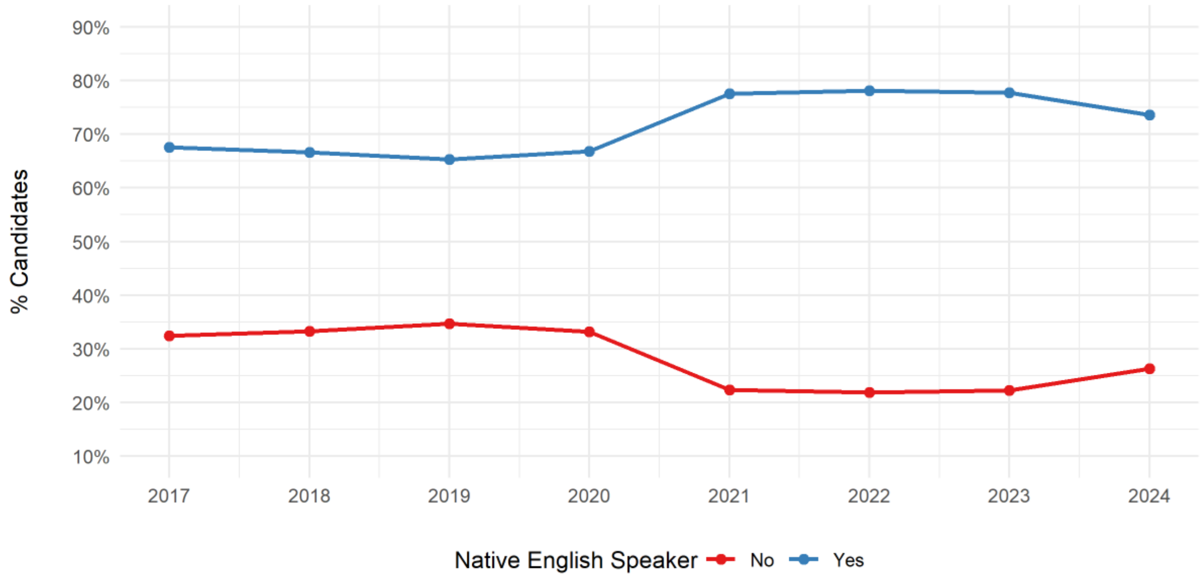
| Subtest | *F*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 108.29 | 2 | < 0.01 |
| DM | 57.38 | 2 | < 0.01 |
| QR | 90.82 | 2 | < 0.01 |
| AR | 163.19 | 2 | < 0.01 |

| Subtest | *F*-Statistic | *df* | *p* Value |
|---|---|---|---|
| Total Cognitive | 134.99 | 2 | < 0.01 |
| SJT | 1,218.60 | 2 | < 0.01 |

### 4.5.8 First Language

In 2024, 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 15). Between 2023 and 2024, 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 15. Count of Language 2017–2024



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 36 below.

Table 36. Scaled Scores by Language and Country of Residence

| Subtest | Country of Residence | First Language | *N* | % of *N* | Mean | *SD* |
|---|---|---|---|---|---|---|
| VR | UK | English | 23,171 | 61% | 611.30 | 73.64 |
| | | Other | 5,991 | 16% | 576.43 | 74.51 |
| | non-UK | English | 4,747 | 13% | 615.82 | 87.16 |
| | | Other | 4,004 | 11% | 559.75 | 77.66 |
| DM | UK | English | 23,171 | 61% | 630.49 | 86.30 |

| Subtest | Country of Residence | First Language | *N* | % of *N* | Mean | *SD* |
|---|---|---|---|---|---|---|
| | | Other | 5,991 | 16% | 592.29 | 89.74 |
| | non-UK | English | 4,747 | 13% | 635.25 | 96.44 |
| | | Other | 4,004 | 11% | 582.61 | 93.85 |
| QR | UK | English | 23,171 | 61% | 656.99 | 91.14 |
| | | Other | 5,991 | 16% | 627.97 | 93.26 |
| | non-UK | English | 4,747 | 13% | 661.75 | 107.69 |
| | | Other | 4,004 | 11% | 616.94 | 102.15 |
| AR | UK | English | 23,171 | 61% | 662.83 | 100.46 |
| | | Other | 5,991 | 16% | 642.39 | 101.21 |
| | non-UK | English | 4,747 | 13% | 647.76 | 106.52 |
| | | Other | 4,004 | 11% | 622.81 | 106.65 |
| Total Cognitive | UK | English | 23,171 | 61% | 2,561.60 | 285.25 |
| | | Other | 5,991 | 16% | 2,439.09 | 300.15 |
| | non-UK | English | 4,747 | 13% | 2,560.58 | 334.92 |
| | | Other | 4,004 | 11% | 2,382.11 | 321.34 |
| SJT | UK | English | 23,171 | 61% | 602.14 | 62.63 |
| | | Other | 5,991 | 16% | 582.30 | 76.50 |
| | non-UK | English | 4,747 | 13% | 576.92 | 76.50 |
| | | Other | 4,004 | 11% | 529.68 | 96.07 |

In line with the other demographic categories, a test was carried out to understand whether the differences observed in Table 36 can be considered statistically significant. Table 37 shows that that such differences between the two groups are unlikely to have occurred by chance.

Table 37. Language *t*-Test

| Subtest | *t*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 47.67 | 37,911 | < 0.01 |
| DM | 41.33 | 37,911 | < 0.01 |
| QR | 30.95 | 37,911 | < 0.01 |
| AR | 21.58 | 37,911 | < 0.01 |
| Total Cognitive | 41.72 | 37,911 | < 0.01 |
| SJT | 43.29 | 37,911 | < 0.01 |

### 4.5.9 Demographic Interactions and SEN

The way demographic characteristics influence UCAT scores is fairly well known. In 2020, Pearson VUE undertook an analysis of variance to explore the interaction between demographic variables and SEN exams. The demographic variables were found to have a significant influence on scores across all cognitive subtests. Furthermore, statistically significant relationships were identified between SEN 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 shortened version of that analysis of variance was also conducted this year to continue monitoring the differences in the performance between UCAT candidates and UCATSEN candidates. The results are presented in Table 38. After controlling for the effect of the demographic variables (see the note in Table 38), the exam version still accounted for a significant amount of variance in performance, with UCATSEN candidates outperforming UCAT candidates. The largest difference was observed in the AR subtest, while the smallest difference was in the QR subtest, closely followed by the SJT, consistent with findings from 2023. In contrast, the 2022 analysis found the largest difference in QR and the smallest in SJT, which aligns with the most and least speeded subtests of the exam, respectively. This pattern in 2022 led to the hypothesis that the performance advantage for SEN candidates may be positively associated with the speededness of the subtest. However, results from this year and the previous year contradict this hypothesis, as QR and AR—both relatively speeded subtests—show differing patterns of performance differences. The performance differences between UCAT and UCATSEN candidates will continue to be monitored in future years to ensure the fairness of the test for all candidates.

Table 38. Subtest Performance Differences: UCAT and UCATSEN (controlling for demographic variables)

| Subtest | *F* | *p* | $\eta^2$ |
|---|---|---|---|
| VR | 91.91 | <.0001 | 0.0023 |
| DM | 134.71 | <.0001 | 0.0034 |
| QR | 68.31 | <.0001 | 0.0017 |
| AR | 187.66 | <.0001 | 0.0049 |
| SJT | 68.69 | <.0001 | 0.0017 |

*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, $\eta^2$) of these differences across all subtests is less than 0.005 after controlling for the effect of the demographic variables, indicating the effect sizes of the differences are very small. The small effect size suggests that the performance gap is not worryingly large considering the normal variation in participants' performance after accounting for the differences in candidates' demographic composition.

# 5. Exam Timing Analysis

The section time for each candidate is calculated by summing the item and review time for each item and candidate. Table 39 shows the exam timing for each version of the UCAT.

Table 39. Subtest Section Timing: Non-SEN and SEN

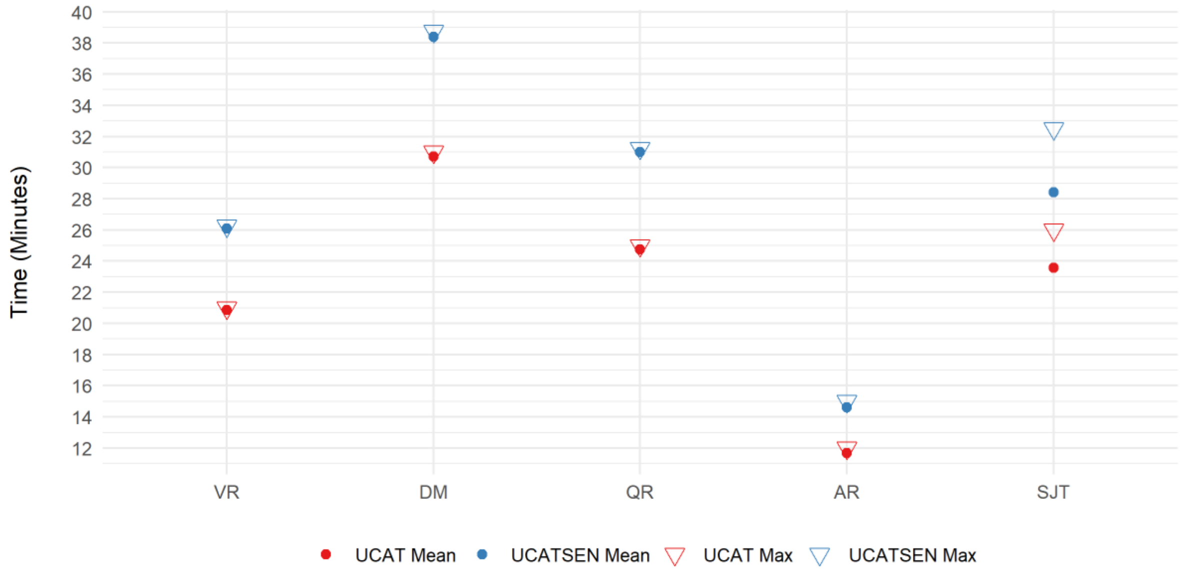
| Statistic | Subtest | UCAT (35565) | UCATSEN (1454) | UCATSEN SA (590) | UCATSEN 50 (64) | UCATSEN 50SA (41) | UCATSA (196) |
|---|---|---|---|---|---|---|---|
| Mean | VR | 00:20:52 | 00:26:05 | 00:26:03 | 00:31:14 | 00:31:21 | 00:20:50 |
| | DM | 00:30:43 | 00:38:24 | 00:38:18 | 00:45:36 | 00:46:13 | 00:30:43 |
| | QR | 00:24:44 | 00:31:00 | 00:30:57 | 00:36:57 | 00:37:13 | 00:24:47 |
| | AR | 00:11:40 | 00:14:36 | 00:14:33 | 00:17:06 | 00:17:32 | 00:11:41 |
| | SJT | 00:23:34 | 00:28:24 | 00:27:16 | 00:31:59 | 00:31:14 | 00:23:13 |
| *SD* | VR | 00:00:28 | 00:00:26 | 00:01:07 | 00:00:55 | 00:00:19 | 00:00:24 |
| | DM | 00:01:03 | 00:01:07 | 00:02:05 | 00:02:24 | 00:00:52 | 00:00:41 |
| | QR | 00:01:12 | 00:00:53 | 00:01:34 | 00:01:42 | 00:00:56 | 00:00:38 |
| | AR | 00:00:47 | 00:00:55 | 00:01:03 | 00:02:13 | 00:01:00 | 00:00:45 |
| | SJT | 00:03:29 | 00:05:10 | 00:05:54 | 00:07:47 | 00:08:19 | 00:03:42 |
| Min | VR | 00:01:51 | 00:19:07 | 00:01:05 | 00:24:17 | 00:29:44 | 00:16:54 |
| | DM | 00:01:51 | 00:22:57 | 00:06:13 | 00:34:09 | 00:41:04 | 00:24:41 |
| | QR | 00:00:39 | 00:04:22 | 00:04:57 | 00:26:17 | 00:31:42 | 00:18:31 |
| | AR | 00:00:45 | 00:04:27 | 00:01:59 | 00:05:52 | 00:13:13 | 00:06:07 |
| | SJT | 00:01:14 | 00:01:38 | 00:10:31 | 00:14:48 | 00:13:34 | 00:08:53 |
| Max | VR | 00:21:00 | 00:26:15 | 00:26:15 | 00:31:30 | 00:31:30 | 00:21:00 |
| | DM | 00:31:00 | 00:38:45 | 00:38:45 | 00:46:30 | 00:46:30 | 00:31:00 |
| | QR | 00:25:00 | 00:31:15 | 00:31:15 | 00:37:30 | 00:37:30 | 00:25:00 |
| | AR | 00:12:00 | 00:15:00 | 00:15:00 | 00:18:00 | 00:18:00 | 00:12:00 |
| | SJT | 00:26:00 | 00:32:30 | 00:32:30 | 00:39:00 | 00:39:00 | 00:26:00 |

*Note.* UCATSEN100 and 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.

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 39. 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 16. For both UCAT and UCATSEN, the difference between the average time used and the maximum time allowed is almost negligible for VR and QR. The difference is slightly more noticeable for DM and AR and becomes quite clear for the SJT.

Figure 16. Mean and Maximum Time for UCAT and UCATSEN



Test timing is examined in more detail in Table 40. It shows that the most speeded non-SEN subtests are VR and QR, where 86% and 88% of candidates respectively reached all the items and between 5% to 7% of candidates did not reach five or more items. The SJT is the least speeded in all exam versions.

Table 40. Subtest Section Timing: Non-SEN and SEN UCAT Incomplete Tests

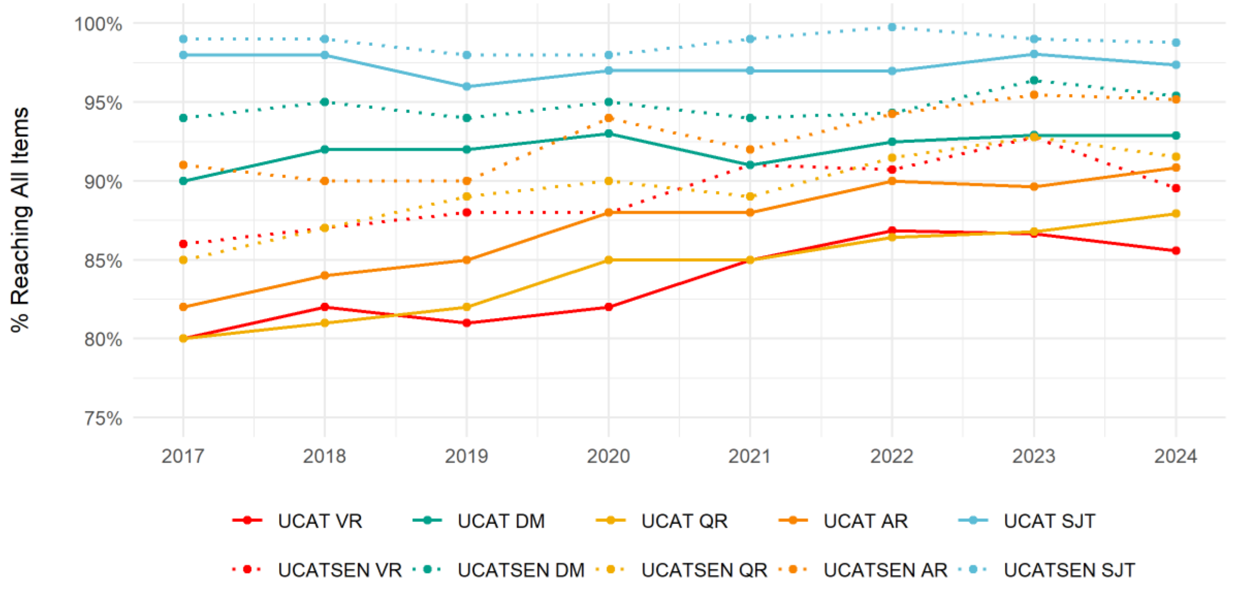
| Exam | Subtest | Reached All Items *N* | Reached All Items % | Five or More Items Unreached *N* | Five or More Items Unreached % | Mean Number of Unreached Items for Incomplete Tests Only |
|---|---|---|---|---|---|---|
| UCAT | VR | 30,435 | 86% | 2,617 | 7% | 6.86 (5130) |
| | DM | 33,028 | 93% | 716 | 2% | 3.65 (2537) |
| | QR | 31,277 | 88% | 1,952 | 5% | 5.74 (4288) |
| | AR | 32,307 | 91% | 1,489 | 4% | 6.38 (3258) |
| | SJT | 34,627 | 97% | 170 | 0% | 3.58 (938) |
| UCATSEN | VR | 1,302 | 90% | 68 | 5% | 5.6 (152) |
| | DM | 1,387 | 95% | 10 | 1% | 2.88 (67) |
| | QR | 1,331 | 92% | 44 | 3% | 4.73 (123) |
| | AR | 1,384 | 95% | 27 | 2% | 5.01 (70) |
| | SJT | 1,436 | 99% | 2 | 0% | 2.44 (18) |
| UCATSENSA | VR | 535 | 91% | 22 | 4% | 5.27 (55) |
| | DM | 571 | 97% | 5 | 1% | 3.47 (19) |
| | QR | 549 | 93% | 15 | 3% | 5.61 (41) |
| | AR | 556 | 94% | 9 | 2% | 4.79 (34) |
| | SJT | 588 | 100% | 0 | 0% | 1.5 (2) |
| UCATSEN50 | VR | 60 | 94% | 2 | 3% | 3.5 (4) |
| | DM | 62 | 97% | 0 | 0% | 1 (2) |
| | QR | 62 | 97% | 1 | 2% | 9.5 (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 |
|---|---|---|---|---|---|---|
| | AR | 63 | 98% | 1 | 2% | 5 (1) |
| | SJT | 64 | 100% | 0 | 0% | - |
| UCAT50SA | VR | 34 | 83% | 3 | 7% | 4 (7) |
| | DM | 40 | 98% | 1 | 2% | 8 (1) |
| | QR | 38 | 93% | 1 | 2% | 2.67 (3) |
| | AR | 38 | 93% | 0 | 0% | 2.33 (3) |
| | SJT | 41 | 100% | 0 | 0% | - |
| UCATSA | VR | 181 | 92% | 12 | 6% | 5.93 (15) |
| | DM | 187 | 95% | 1 | 1% | 3 (9) |
| | QR | 182 | 93% | 6 | 3% | 3.93 (14) |
| | AR | 185 | 94% | 6 | 3% | 5.09 (11) |
| | SJT | 195 | 99% | 0 | 0% | 1 (1) |

*Note.* UCATSEN100 and 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 17 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. In 2024, VR showed a slight decrease in the percentage of candidates completing all items compared to 2023, indicating it has become slightly more speeded. However, QR and AR continued to show improvement, with a higher percentage of candidates completing the test.

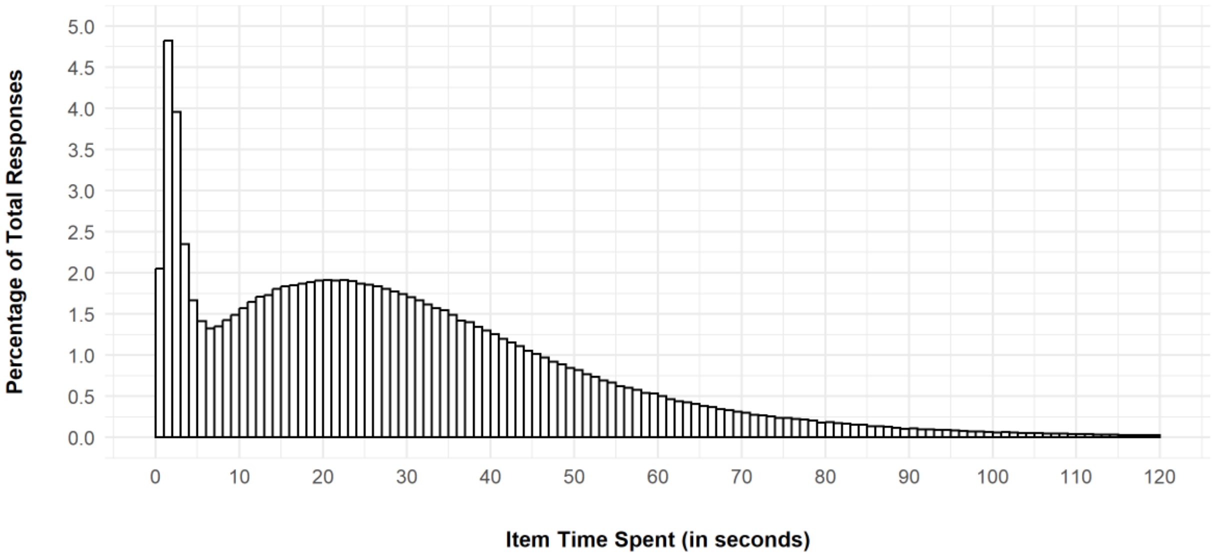
Figure 17. Candidates Reaching All Items 2017–2024



While item timing had long been considered in the form construction for QR and AR, this approach was extended to VR and DM in 2022, helping all subtests become less speeded over time. In 2022, timing adjustments were made to AR and QR to reduce QR's speededness. One minute was removed from AR, along with 5 pretest items, and added to QR without increasing the number of items. Following these changes, a notable increase in the percentage of candidates completing all items was observed, making QR less speeded.

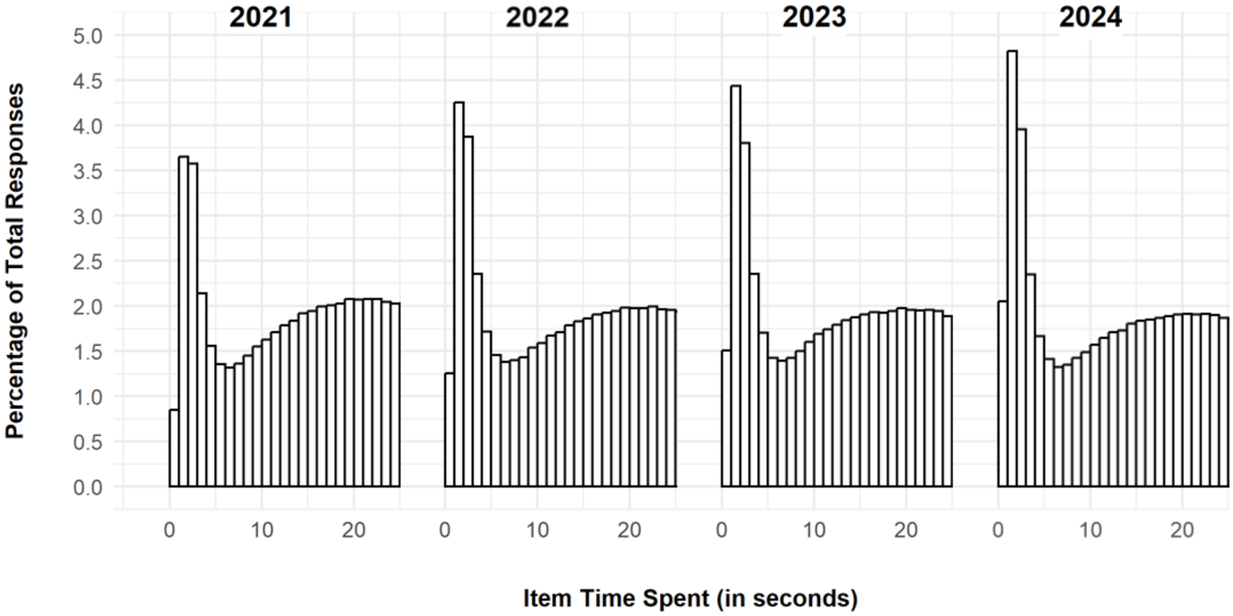
In 2023, the VR item composition was adjusted to improve test discrimination, reducing "True-False-Can't Tell" questions by 10% and increasing "Multiple Choice" questions by 10%, as "Multiple Choice" items are more discriminating. However, these items are slightly longer due to the additional text in the options, potentially making VR marginally more speeded. In 2024, a further adjustment reduced "True-False-Can't Tell" questions by another 10%, with a corresponding increase in "Multiple Choice" questions. As Figure 17 shows, this resulted in a slight decline in the percentage of candidates completing all items, indicating a modest increase in speededness. At the same time, as shown later in the technical report (Figure 29), these changes led to a small increase in the VR subtest's point-biserial, reflecting improved item discrimination.

Figure 18. VR Response Time Distribution – 2024



The factor of guessing has been considered when evaluating speededness since 2022. Figure 18 to Figure 25 illustrate the distribution of item response times for the five 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. By excluding responses shorter than the valley duration, it is possible to filter out most 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.

Figure 19. VR Response Time Distribution – 2021 to 2024



When examining the VR item time distribution from 2021 to 2024, the first peak of the distribution is observed around 2 to 3 seconds. As shown in Figure 19, in 2021, the peak accounted for approximately 3.5% of total responses, with responses made between 0 and 5 seconds comprising about 12% of the total. By 2024, the peak had increased to around 4.8% of total responses, and the proportion of responses made between 0 and 5 seconds had risen to approximately 15%. This indicates that the speededness of the VR subtest has gradually increased over time, with a larger proportion of responses in 2024 being made within 5 seconds, likely representing guessed responses.

Figure 20. DM Response Time Distribution – 2024

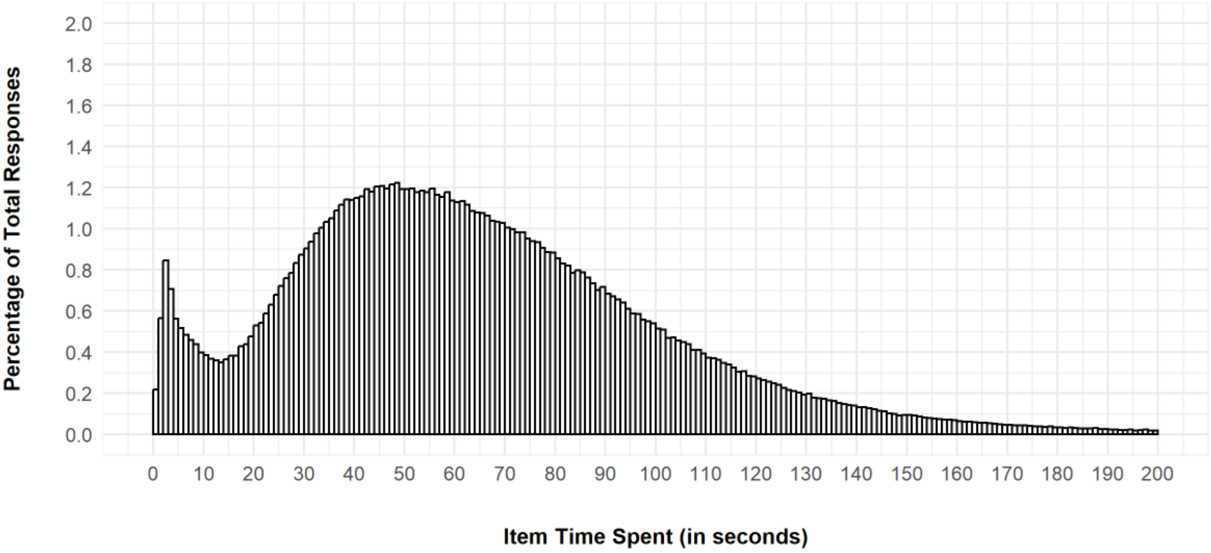


Figure 20 illustrates the response time distribution for the DM subtest in 2024. Unlike VR and QR, the height of the first peak, representing responses made between 2 to 3 seconds, is relatively small, accounting for approximately 0.85% of total responses. The majority of responses fall within the main 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 21 shows the response time distributions for DM from 2021 to 2024. While there are minor fluctuations and changes across the years, the distributions remain largely consistent over time.

Figure 21. DM Response Time Distribution – 2021 to 2024

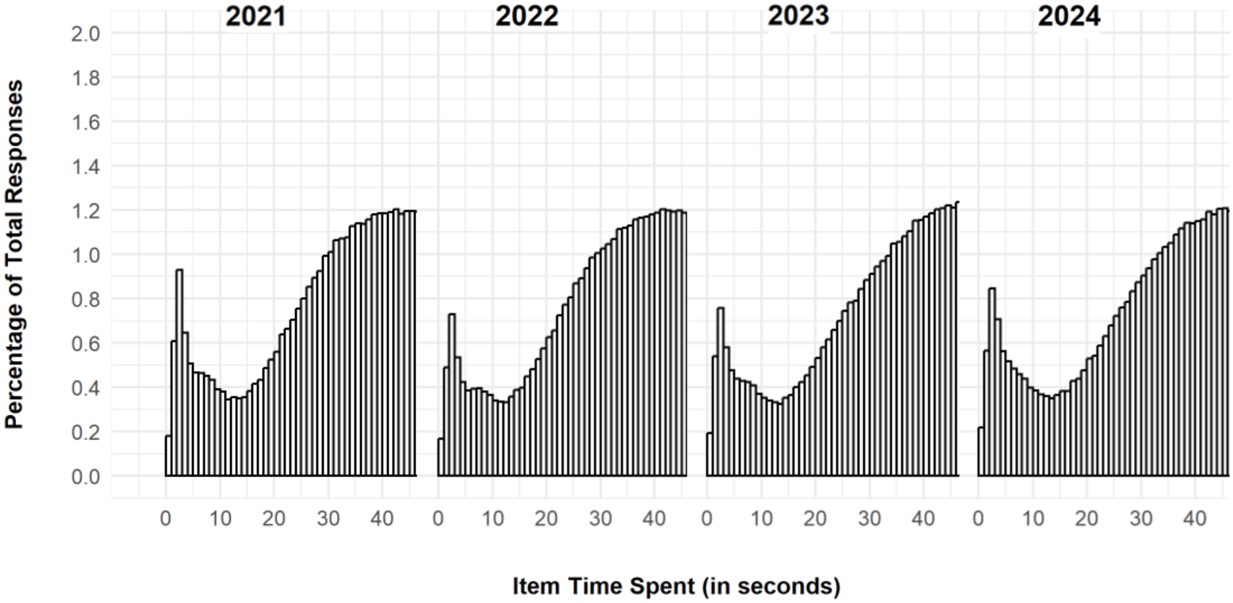
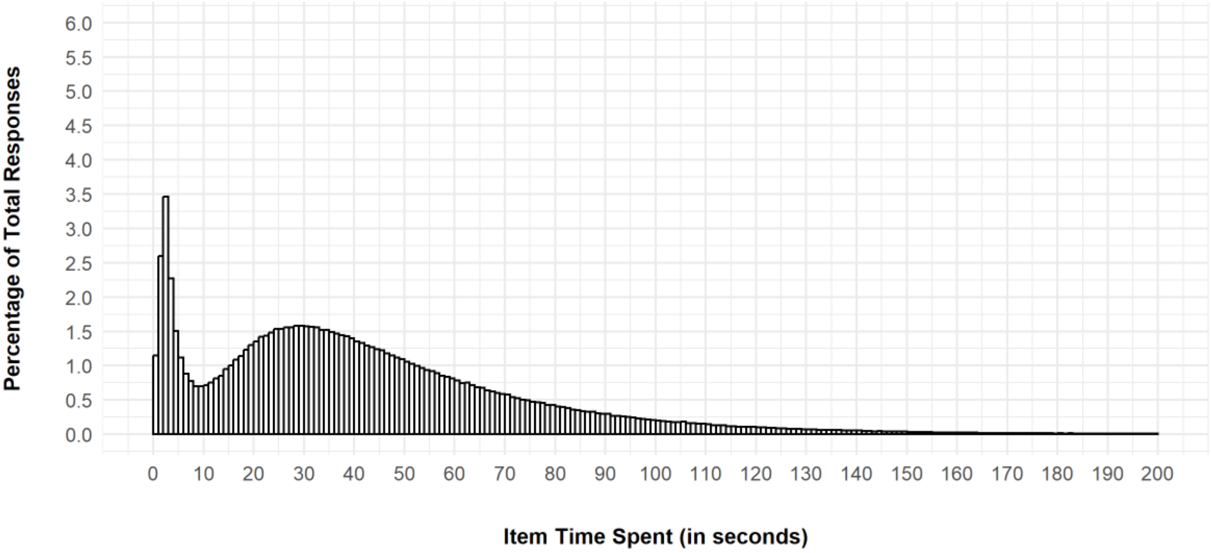
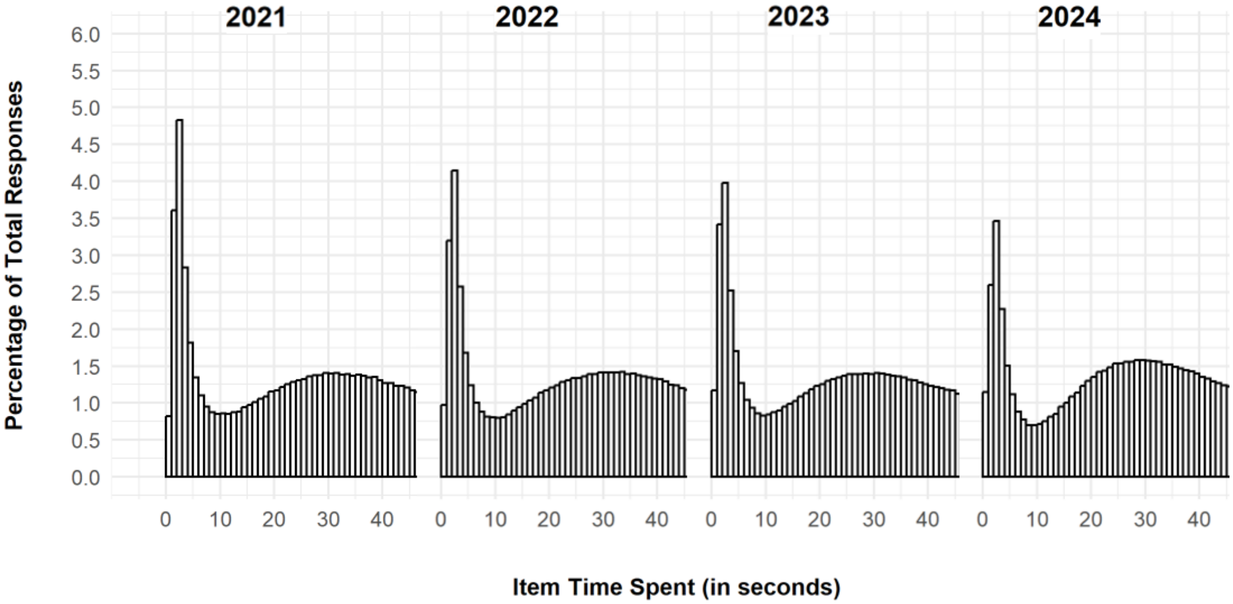


Figure 22. QR Response Time Distribution – 2024



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 22 shows the QR response time distribution in 2024. The first peak, representing guessed responses between 2 to 3 seconds, accounts for approximately 3.5% of total responses, while 11% of responses were made in under 5 seconds. Figure 23 illustrates the QR distributions from 2021 to 2024, showing an opposite trend to VR. The first peak (guessed responses) decreased from around 4.8% in 2021 to 3.5% in 2024, and the proportion of responses below 5 seconds dropped from about 14% in 2021 to 11% in 2024. This indicates an improvement in the speededness of the QR subtest over time.

Figure 23. QR Response Time Distribution – 2021 to 2024



In contrast to VR, DM, and QR, the AR and SJT subtests display skewed unimodal distributions, as shown in Figure 24 and Figure 25. 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 24. AR Response Time Distribution – 2024

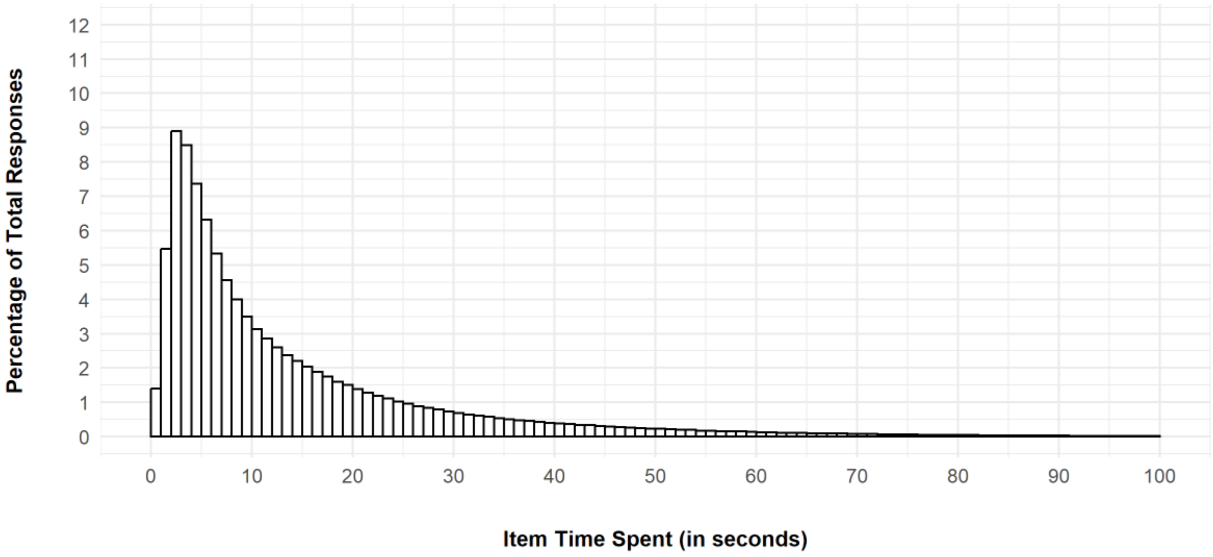
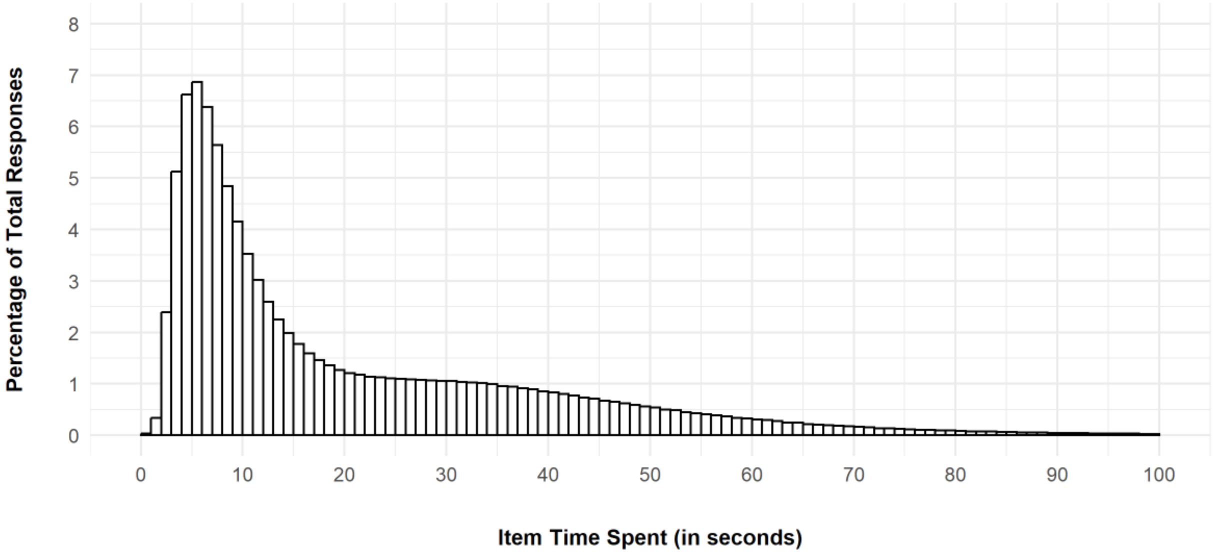


Figure 25. SJT Response Time Distribution – 2024



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 AR and SJT make applying a fixed threshold less effective and may inadvertently exclude a significant number of non-guessed responses. Hence, the similar analysis for AR and the SJT are intentionally omitted in Table 41 to avoid unnecessary confusion.

Using a balanced 5-second exclusion threshold, the proportion of candidates completing all items in VR, DM, and QR without guessing dropped significantly to 13%, 64%, and 27%, respectively. This highlights that only a small proportion of candidates were able to complete these subtests within the allotted time without resorting to guessing. However, on average, candidates reached 83%, 96%, and 87% of the items in VR, DM, and QR, respectively. This suggests that while many candidates did not finish every item without guessing, they were generally able to attempt most items. Regardless of the guessing exclusion, VR and QR remained the most speeded subtests, with VR being slightly more speeded than QR. The overall results showed a slight decrease in completion rates for VR after excluding guessed responses, while QR showed an increase. Notably, 27% of candidates completed QR without guessing, up from 20% in 2023, indicating a noticeable improvement in QR speededness. Conversely, VR saw a slight decline, reflecting a worsening in its speededness.

Table 41. 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 | 86% | 98% | 100% | 100% | 100% |
| | Excluding responses ≤ 1s | 64% | 96% | 95% | 100% | 100% |
| | Excluding responses ≤ 5s | 13% | 83% | 75% | 86% | 95% |
| | Excluding responses ≤ 10s | 1% | 76% | 68% | 77% | 86% |
| DM | All responses included | 93% | 99% | 100% | 100% | 100% |
| | Excluding responses ≤ 1s | 90% | 99% | 100% | 100% | 100% |
| | Excluding responses ≤ 5s | 64% | 96% | 97% | 100% | 100% |
| | Excluding responses ≤ 10s | 48% | 94% | 90% | 97% | 100% |
| QR | All responses included | 88% | 98% | 100% | 100% | 100% |
| | Excluding responses ≤ 1s | 76% | 97% | 100% | 100% | 100% |
| | Excluding responses ≤ 5s | 27% | 87% | 81% | 92% | 100% |
| | Excluding responses ≤ 10s | 15% | 83% | 75% | 86% | 97% |
| AR and SJT results are intentionally omitted to avoid confusion | | | | | | |

# 6. Test Form Analysis

Table 42 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 42. Candidates by Form

| Form | Candidates |
|---|---|
| Form 1 | 7,848 |
| Form 2 | 7,843 |
| Form 3 | 7,980 |
| Form 4 | 7,187 |
| Form 5 | 7,055 |

Table 43 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 43. Cognitive Raw Score Test Statistics

| Subtest | Form | Mean | *SD* | Min | Max | Alpha | *SEM* |
|---|---|---|---|---|---|---|---|
| VR<br>(40 items) | Form 1 | 21.28 | 6.04 | 3 | 39 | 0.76 | 2.96 |
| | Form 2 | 21.56 | 6.01 | 2 | 39 | 0.76 | 2.94 |
| | Form 3 | 21.66 | 6.13 | 1 | 40 | 0.78 | 2.88 |
| | Form 4 | 21.41 | 5.79 | 2 | 39 | 0.74 | 2.95 |
| | Form 5 | 21.3 | 6.24 | 2 | 39 | 0.78 | 2.93 |
| DM<br>(26 items; 34 score points) | Form 1 | 19.52 | 5.59 | 1 | 33 | 0.75 | 2.8 |
| | Form 2 | 17.9 | 6.1 | 2 | 34 | 0.78 | 2.86 |
| | Form 3 | 17.85 | 5.71 | 2 | 33 | 0.75 | 2.86 |
| | Form 4 | 18.38 | 5.42 | 2 | 33 | 0.73 | 2.82 |
| | Form 5 | 17.98 | 5.93 | 2 | 33 | 0.77 | 2.84 |
| QR<br>(32 items) | Form 1 | 19.83 | 6.34 | 0 | 32 | 0.86 | 2.37 |
| | Form 2 | 19.68 | 6.07 | 0 | 32 | 0.84 | 2.43 |
| | Form 3 | 20.24 | 6.27 | 2 | 32 | 0.85 | 2.43 |
| | Form 4 | 20.31 | 6.28 | 2 | 32 | 0.85 | 2.43 |
| | Form 5 | 20.43 | 7.07 | 1 | 32 | 0.89 | 2.34 |
| AR<br>(50 items) | Form 1 | 34.11 | 8.04 | 3 | 50 | 0.86 | 3.01 |
| | Form 2 | 33.87 | 8.32 | 4 | 50 | 0.87 | 3 |
| | Form 3 | 33.52 | 8.28 | 4 | 50 | 0.86 | 3.1 |
| | Form 4 | 33.48 | 7.93 | 2 | 50 | 0.86 | 2.97 |
| | Form 5 | 32.69 | 7.63 | 5 | 50 | 0.84 | 3.05 |

Table 43 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.

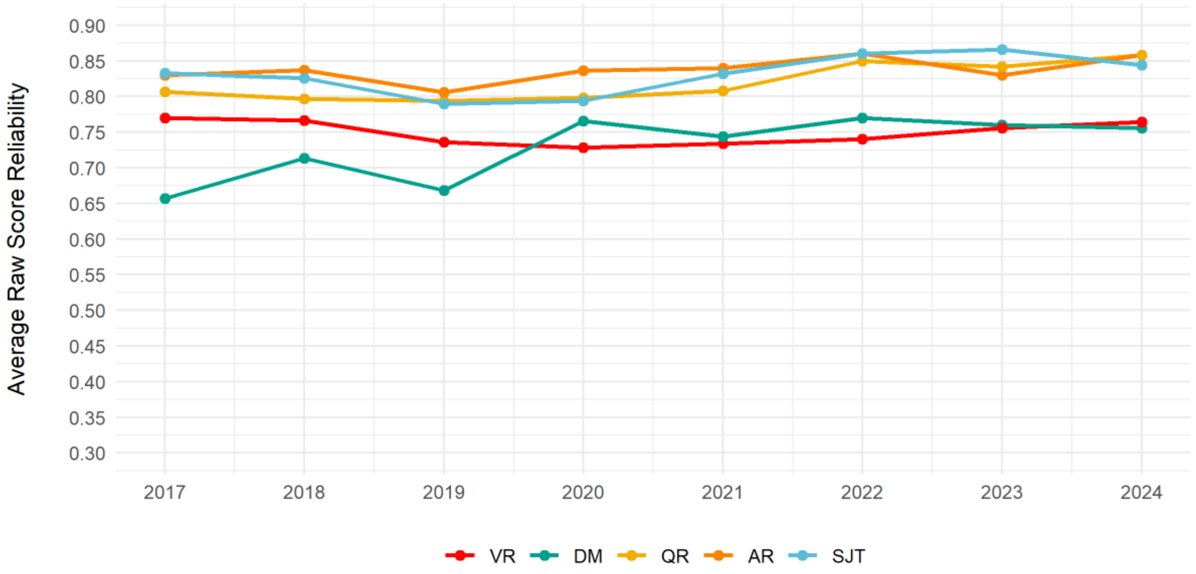
The SJT is analysed in a similar way to the cognitive sections above; however, because the maximum raw score available on the SJT can change year on year, an additional column called mean percent raw score is added (Table 44). Similar to the cognitive results, the reliability is adequately high and the *SEM* adequately low for the SJT.

Table 44. SJT Raw Score Test Statistics (252 score points)

| Form | Mean | *SD* | Min | Max | Mean Percent Raw Score | Alpha | *SEM* |
|---|---|---|---|---|---|---|---|
| Form 1 | 191.29 | 21.23 | 66 | 236 | 75.91% | 0.84 | 8.49 |
| Form 2 | 193.21 | 20.50 | 61 | 236 | 76.67% | 0.83 | 8.45 |
| Form 3 | 194.46 | 20.83 | 66 | 239 | 77.17% | 0.85 | 8.07 |
| Form 4 | 190.91 | 22.22 | 48 | 234 | 75.76% | 0.86 | 8.31 |
| Form 5 | 193.22 | 21.14 | 51 | 235 | 76.68% | 0.84 | 8.46 |

Figure 26 shows the mean Cronbach's alpha for each subtest in each form since 2017. Note that prior to 2019, it is the mean of three forms, whereas since 2019, it is the mean of five forms. DM has become more reliable since its launch in 2017, and the reliability of VR has slightly dropped but remained consistent since 2020, with a small improvement in 2024. The reliability of both QR and the SJT has continued to improve this year.

Figure 26. Raw Score Reliability 2017–2024



Raw scores are scaled and reported as scaled scores. The summary statistics for scaled scores on each form are presented below in Table 45. 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 45 also includes the results for the SJT.

Table 45. Cognitive Scaled Score Test Statistics

| Subtest | Form | Mean | *SD* | Min | Max | Reliability | *SEM* |
|---|---|---|---|---|---|---|---|
| VR | Form 1 | 598.5 | 77.11 | 310 | 880 | 0.75 | 38.56 |
| | Form 2 | 603.12 | 77.71 | 300 | 880 | 0.75 | 38.85 |
| | Form 3 | 603.86 | 81.63 | 300 | 900 | 0.77 | 39.15 |
| | Form 4 | 599.36 | 74.29 | 300 | 880 | 0.73 | 38.6 |
| | Form 5 | 599.4 | 80.76 | 300 | 880 | 0.77 | 38.73 |
| DM | Form 1 | 625.95 | 91.17 | 300 | 890 | 0.76 | 44.66 |
| | Form 2 | 619.36 | 95.33 | 300 | 900 | 0.79 | 43.69 |
| | Form 3 | 614.5 | 88.04 | 300 | 890 | 0.75 | 44.02 |
| | Form 4 | 623.24 | 84.85 | 300 | 890 | 0.74 | 43.27 |
| | Form 5 | 616.99 | 94.74 | 300 | 890 | 0.78 | 44.44 |
| QR | Form 1 | 643.33 | 90.97 | 300 | 900 | 0.82 | 38.6 |
| | Form 2 | 643.83 | 91.01 | 300 | 900 | 0.81 | 39.67 |
| | Form 3 | 649.59 | 93.97 | 350 | 900 | 0.81 | 40.96 |
| | Form 4 | 650.53 | 92.51 | 350 | 900 | 0.81 | 40.32 |
| | Form 5 | 657.57 | 111.41 | 300 | 900 | 0.83 | 45.94 |
| AR | Form 1 | 659.9 | 103.56 | 300 | 900 | 0.84 | 41.42 |
| | Form 2 | 659.21 | 109.29 | 300 | 900 | 0.84 | 43.72 |
| | Form 3 | 655.31 | 105.94 | 300 | 900 | 0.84 | 42.38 |
| | Form 4 | 650.97 | 99.41 | 300 | 900 | 0.83 | 40.99 |
| | Form 5 | 640.48 | 93 | 300 | 900 | 0.82 | 39.46 |
| Total Cognitive | Form 1 | 2,527.68 | 296.23 | 1,410 | 3,500 | 0.92 | 83.79 |
| | Form 2 | 2,525.52 | 309.08 | 1,440 | 3,540 | 0.93 | 81.77 |
| | Form 3 | 2,523.25 | 310.32 | 1,430 | 3,550 | 0.93 | 82.1 |
| | Form 4 | 2,524.1 | 288.51 | 1,400 | 3,470 | 0.92 | 81.6 |
| | Form 5 | 2,514.44 | 321.38 | 1,200 | 3,560 | 0.93 | 85.03 |
| SJT | Form 1 | 582.98 | 74.46 | 300 | 742 | 0.84 | 29.78 |
| | Form 2 | 590.25 | 73.63 | 300 | 746 | 0.83 | 30.36 |
| | Form 3 | 594.49 | 74.02 | 300 | 754 | 0.85 | 28.67 |
| | Form 4 | 581.95 | 76.81 | 300 | 733 | 0.86 | 28.74 |
| | Form 5 | 590.96 | 72.09 | 300 | 734 | 0.84 | 28.84 |

# 7. Item Analysis

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

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

## 7.1 Cognitive Item Analysis

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

- Point biserial: the degree to which a test item discriminated between strong and weak candidates. For operational items, it must be greater than 0.1 for the item to remain in the bank. For pretest items, it must be greater than 0.05.
- *p* Value: the proportion of candidates who answered the item correctly—the item difficulty. This must be between 0.1 and 0.95 for the item to remain in the bank.
- IRT *b*: the difficulty parameter from the item response theory 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 46 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 46. Cognitive Items Passing the Quality Criteria

| | | VR | | DM | | QR | | AR | |
|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational Scored | Pass | 200 | 100% | 129 | 99% | 160 | 100% | 250 | 100% |
| | Fail | 0 | 0% | 1 | 1% | 0 | 0% | 0 | 0% |
| | *p* < 10 or > 95 | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | pBis <= 0.1 | 0 | 0% | 1 | 1% | 0 | 0% | 0 | 0% |
| | \|*b*\| >= 3 | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest Unscored | Pass | 277 | 98% | 247 | 99% | 295 | 100% | N/A | N/A |
| | Fail | 6 | 2% | 2 | 1% | 1 | 0% | N/A | N/A |
| | *p* < 10 or > 95 | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A |
| | pBis <= 0.05 | 6 | 2% | 2 | 1% | 1 | 0% | N/A | N/A |
| | \|*b*\| >= 3 | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A |

Consistent with previous years, only a very small number of operational items failed the analysis. This year, only one DM operational item failed due to insufficient discrimination. For pretest items, a few failures were observed in the VR, DM, and QR subtests, primarily due to low performance. Since 2022, there have been no pretest items for AR. Figure 27 and Figure 28 show that both operational and pretest item pass rates have been improving over time, with excellent overall pass rates.

Figure 27. Proportion of Operational Items Failing Analysis 2017–2024

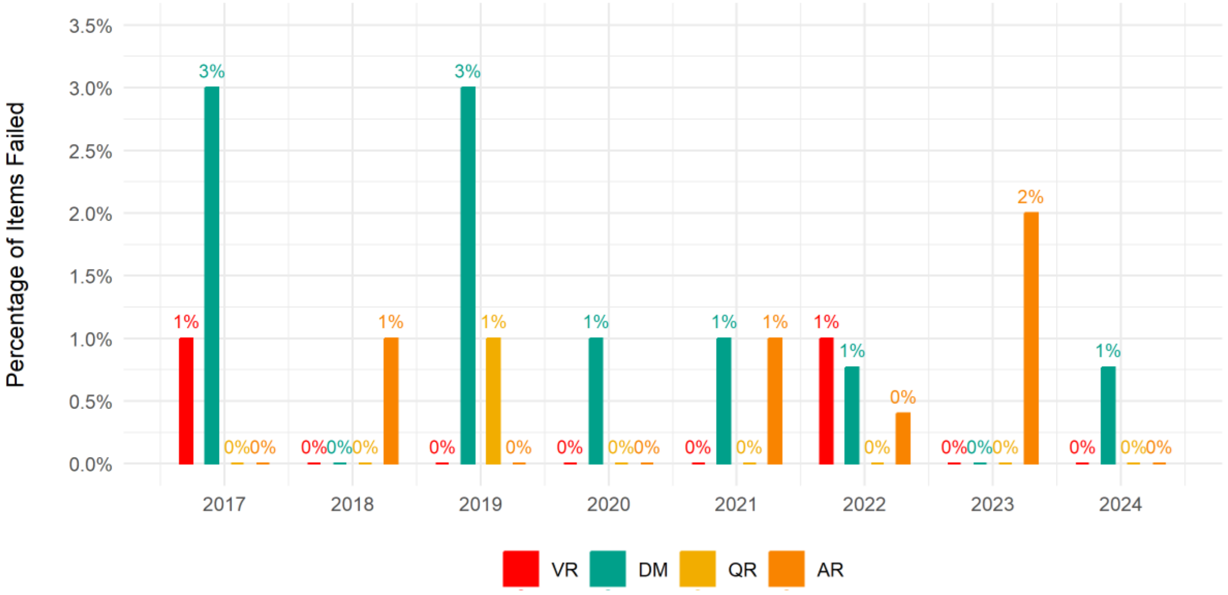


Figure 28. Proportion of Pretest Items Failing Analysis 2017–2024

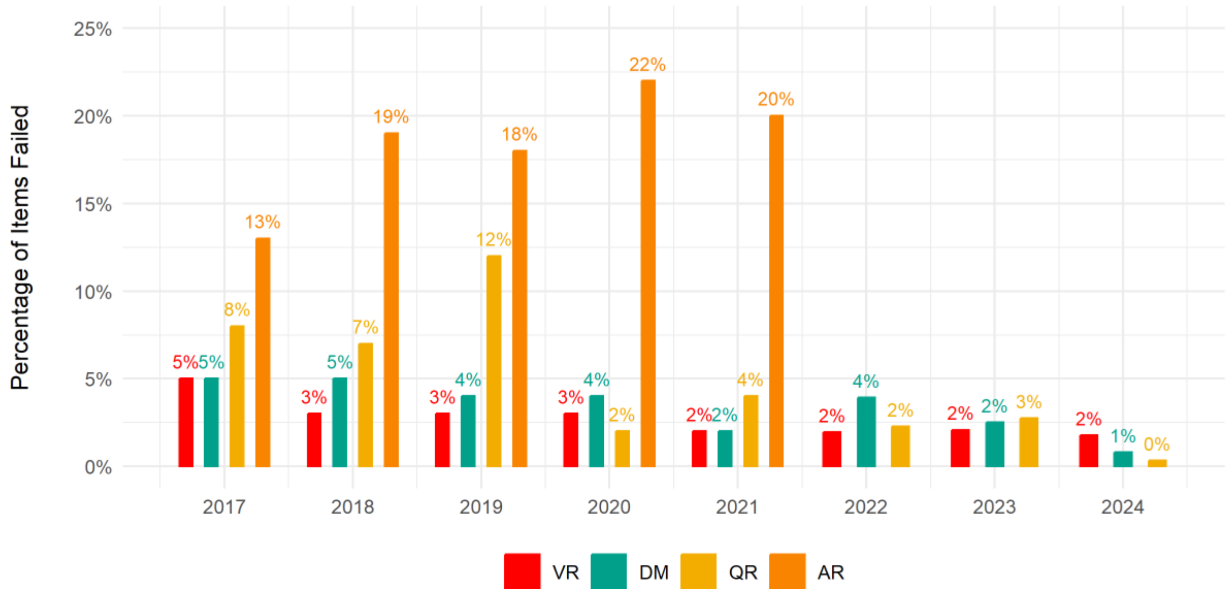


Table 47 shows a summary of the point biserial values. The maximum point biserial is 1, and higher values are better because they indicate that an item can discriminate well between strong and weak candidates. Given that the unscored items have not been tested before, it is expected that those items, on average, will discriminate less well than the scored items, and that is the case across all the cognitive subtests.

Table 47. Discrimination Summary Statistics

| Scored/Unscored | Subtest | *N* Items | Mean pBis | *SD* pBis | Min pBis | Max pBis |
|---|---|---|---|---|---|---|
| Operational (Scored) | VR | 200 | 0.29 | 0.05 | 0.14 | 0.43 |
| | DM | 130 | 0.36 | 0.11 | 0.05 | 0.64 |
| | QR | 160 | 0.41 | 0.07 | 0.15 | 0.58 |
| | AR | 250 | 0.34 | 0.07 | 0.16 | 0.51 |
| Pretest (Unscored) | VR | 283 | 0.28 | 0.09 | -0.03 | 0.43 |
| | DM | 249 | 0.34 | 0.12 | -0.04 | 0.62 |
| | QR | 296 | 0.39 | 0.09 | 0.02 | 0.60 |
| | AR | N/A | N/A | N/A | N/A | N/A |

Historically, the point biserial values for scored items have been high and stable, while those for unscored items have been lower and less consistent, as shown in Figure 29. The point biserial for operational items has remained relatively stable, but pretest items have shown a noticeable increase this year, with QR pretest items exhibiting a particularly large improvement. This indicates that the quality of pretest items has improved over time.

Figure 29. Point biserial 2017–2024

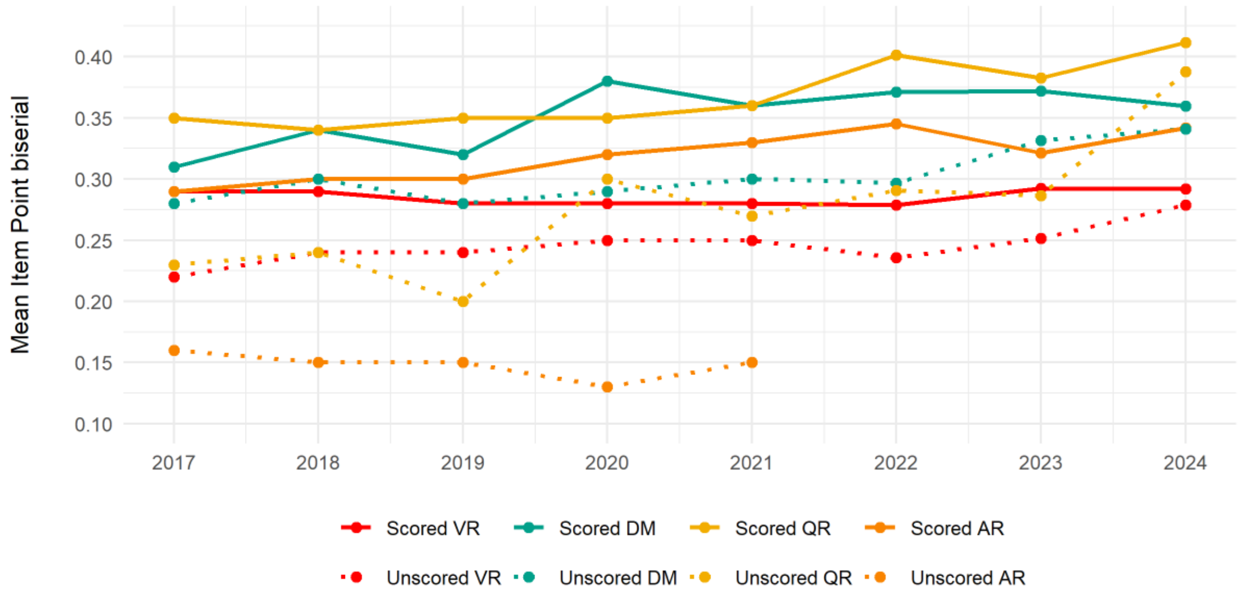


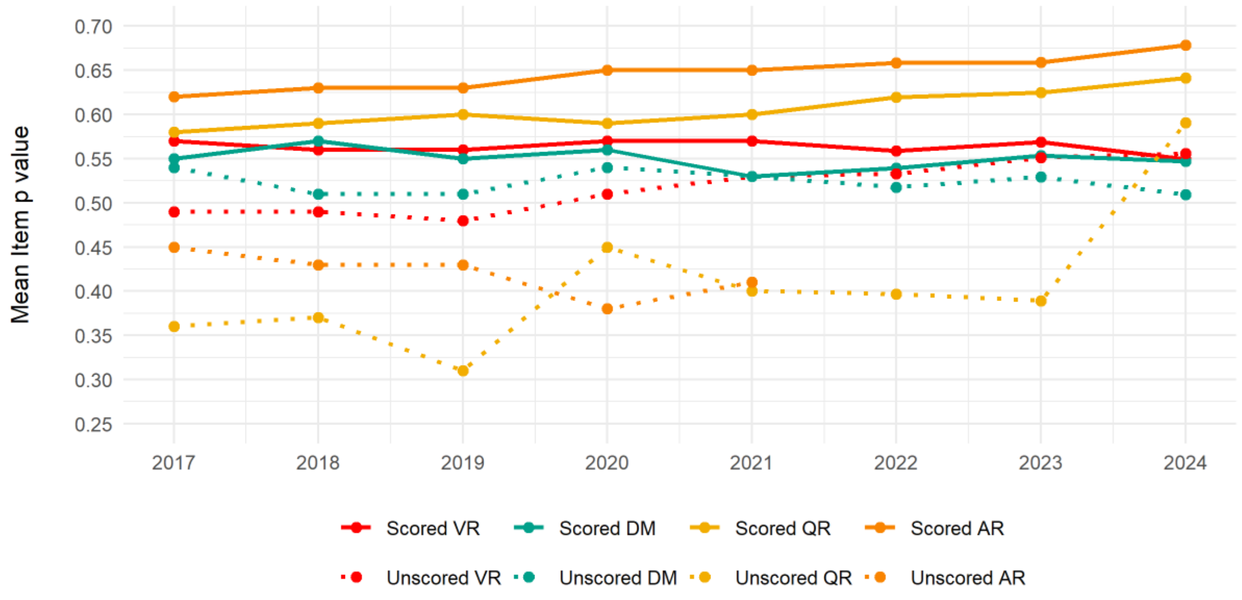
Table 48 summarises the *p* values for the cognitive subtests. *p* values represent the proportion of candidates who answered an item correctly, with higher values indicating easier items and lower values indicating harder items. Among the operational items, VR and DM items were the most difficult on average for 2024 candidates, while AR items were the easiest. For the pretest pools, items in DM and QR were somewhat more difficult than the operational items, whereas VR items were of similar difficulty.

Table 48. *p* Value Summary Statistics

| Scored/Unscored | Subtest | *N* Items | Mean *p* | *SD p* | Min *p* | Max *p* |
|---|---|---|---|---|---|---|
| Operational (Scored) | VR | 200 | 0.55 | 0.13 | 0.23 | 0.86 |
| | DM | 130 | 0.55 | 0.15 | 0.19 | 0.89 |
| | QR | 160 | 0.64 | 0.14 | 0.26 | 0.88 |
| | AR | 250 | 0.68 | 0.13 | 0.26 | 0.94 |
| Pretest (Unscored) | VR | 283 | 0.56 | 0.16 | 0.12 | 0.89 |
| | DM | 249 | 0.51 | 0.18 | 0.13 | 0.93 |
| | QR | 296 | 0.59 | 0.19 | 0.16 | 0.95 |
| | AR | N/A | N/A | N/A | N/A | N/A |

Since 2017, pretesting has been effective in identifying items that are too difficult or too easy. Figure 30 illustrates that items in the pretest pools are typically more difficult, on average, than the operational items. It is important to note that the subtests are equated year-on-year, ensuring that changes in the difficulty of individual items do not affect the ability level required for candidates to achieve a given scaled score. This year, QR pretest items showed a notable increase in *p* value, indicating that the items written for QR in 2024 are not only more discriminating but also easier on average.

Figure 30. *p* Value 2017–2024



The VR subtest consists of four-option multiple-choice items and three-option true/false/can't tell items. Table 49 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 49. 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 | 0.30 | 0.05 | 0.55 | 0.13 |
| | True/False/Can't Tell | 40 | 0.26 | 0.05 | 0.57 | 0.13 |
| Pretest (Unscored) | Multiple Choice | 251 | 0.28 | 0.09 | 0.56 | 0.16 |
| | True/False/Can't Tell | 32 | 0.23 | 0.08 | 0.55 | 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 50.

Table 50. 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 | 40 | 0.46 | 0.09 | 0.53 | 0.15 |
| | Multiple Choice | 90 | 0.31 | 0.09 | 0.56 | 0.14 |
| Pretest (Unscored) | Drag and Drop | 85 | 0.41 | 0.13 | 0.43 | 0.16 |
| | Multiple Choice | 164 | 0.30 | 0.09 | 0.55 | 0.18 |

In addition to different response types, the DM subtest also contains different item types. Among the drag-and-drop items, interpreting information items are more difficult than syllogism items but the latter discriminate slightly better than the former, as presented in Table 51. For the multiple-choice items, the items on statistical reasoning and Venn diagrams are the most discriminating. Statistical reasoning was found to be the most difficult item type in DM, while Venn diagrams were found to be the easiest.

Table 51. 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 | 20 | 0.46 | 0.10 | 0.49 | 0.11 |
| | | Syllogisms | 20 | 0.47 | 0.09 | 0.56 | 0.18 |
| | Multiple Choice | Logical Puzzles | 20 | 0.24 | 0.07 | 0.56 | 0.21 |
| | | Statistical Reasoning | 20 | 0.36 | 0.08 | 0.48 | 0.10 |
| | | Assumptions Recognition | 20 | 0.27 | 0.08 | 0.55 | 0.12 |
| | | Venn Diagrams | 30 | 0.36 | 0.06 | 0.61 | 0.11 |
| Pretest (Unscored) | Drag and Drop | Information Interpretation | 45 | 0.37 | 0.13 | 0.41 | 0.14 |
| | | Syllogisms | 40 | 0.47 | 0.10 | 0.45 | 0.18 |
| | Multiple Choice | Logical Puzzles | 31 | 0.26 | 0.07 | 0.56 | 0.16 |
| | | Statistical Reasoning | 34 | 0.32 | 0.10 | 0.44 | 0.14 |
| | | Assumptions Recognition | 19 | 0.25 | 0.11 | 0.57 | 0.17 |
| | | Venn Diagrams | 80 | 0.33 | 0.08 | 0.59 | 0.19 |

The QR subtest has item sets and standalone items. Each item set contains four items. As with the pretest pool as a whole, the pretest items discriminate less well on average than the ones that have already been pretested prior to appearing in the 2024 exam, as shown in Table 52.

Table 52. 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 | 0.41 | 0.07 | 0.64 | 0.14 |
| | Standalone | 20 | 0.41 | 0.03 | 0.67 | 0.18 |
| Pretest (Unscored) | Item Set | 270 | 0.39 | 0.09 | 0.58 | 0.19 |
| | Standalone | 26 | 0.38 | 0.08 | 0.70 | 0.16 |

The AR subtest consists of four different types. Table 53 below shows that the discrimination of all four item types is similarly strong across the operational items.

Table 53. AR Type Point biserial and *p* Value

| Scored/Unscored | Item Type | *N* Items | Point biserial | | *p* Value | |
|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Mean | *SD* |
| Operational (Scored) | Type 1 | 200 | 0.35 | 0.07 | 0.68 | 0.13 |
| | Type 2 | 10 | 0.29 | 0.07 | 0.74 | 0.13 |
| | Type 3 | 15 | 0.28 | 0.05 | 0.68 | 0.17 |
| | Type 4 | 25 | 0.37 | 0.05 | 0.65 | 0.14 |

### 7.1.1 Item Analysis for SEN

An additional analysis was performed this year to examine whether the items perform differently for exams with accommodations. Overall, the item performances did not show substantial differences between the two set of analyses, with all of the differences being within a third of an *SD* and most of them being within a tenth of an *SD*, as presented in Table 54. 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 54. Item Analysis of UCAT and UCATSEN

| Scored/Unscored | Subtest | Statistics | UCAT | | UCATSEN | |
|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Mean | *SD* |
| Operational (Scored) | VR | *p* Value | 0.55 | 0.13 | 0.59 | 0.14 |
| | | Point biserial | 0.29 | 0.05 | 0.29 | 0.07 |
| | | IRT *b* | -0.21 | 0.66 | -0.20 | 0.68 |
| | DM | Facility | 0.71 | 0.31 | 0.78 | 0.34 |
| | | Point biserial | 0.36 | 0.11 | 0.34 | 0.12 |
| | | IRT *b* | 0.23 | 0.69 | 0.26 | 0.71 |
| | QR | *p* Value | 0.64 | 0.14 | 0.68 | 0.15 |
| | | Point biserial | 0.41 | 0.07 | 0.38 | 0.07 |
| | | IRT *b* | -0.29 | 0.74 | -0.29 | 0.77 |
| | AR | *p* Value | 0.68 | 0.13 | 0.73 | 0.12 |
| | | Point biserial | 0.34 | 0.07 | 0.34 | 0.08 |
| | | IRT *b* | 0.17 | 0.69 | 0.17 | 0.69 |
| Pretest (Unscored) | VR | *p* Value | 0.55 | 0.16 | 0.59 | 0.19 |
| | | Point biserial | 0.28 | 0.09 | 0.28 | 0.22 |
| | | IRT *b* | -0.26 | 0.79 | -0.28 | 1.02 |
| | DM | Facility | 0.65 | 0.28 | 0.71 | 0.33 |
| | | Point biserial | 0.34 | 0.12 | 0.32 | 0.26 |
| | | IRT *b* | 0.38 | 0.91 | 0.45 | 1.01 |
| | QR | *p* Value | 0.59 | 0.19 | 0.63 | 0.22 |

| Scored/Unscored | Subtest | Statistics | UCAT | | UCATSEN | |
|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Mean | *SD* |
| | | Point biserial | 0.39 | 0.09 | 0.32 | 0.24 |
| | | IRT *b* | -0.07 | 1.09 | -0.12 | 1.36 |

## 7.1.2 Comparison of UCAT Item Bank Statistics with UCAT ANZ

The following section is an updated version of the same comparison made in this year's UCAT ANZ technical report with updated item statistics from UCAT 2024. This section presents the performance of test items across the UK and ANZ population of the 2024 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 55 compares the summary statistics for the operational item analysis of the UCAT 2024 and the UCAT ANZ 2024. Across all the subtests, the point biserial summary statistics were similar, with the results from the ANZ population showing slightly higher values, indicating that all operational items discriminated as strongly as expected for the UCAT ANZ population. In terms of the *p* value, which is sample-dependant, the UCAT ANZ population had higher (i.e. easier) average values across subtests. The IRT difficulty, on the other hand, is on a common scale. Table 55 shows that for all subtests, the 2024 UCAT and UCAT ANZ had very similar mean IRT difficulty values, indicating a comparable level of difficulty for both populations.

Table 55. Comparison of Operational Item Statistics: UCAT & UCAT ANZ 2024

| Subtest | Item Statistics | *N* Items | UCAT 2024 | | UCAT ANZ 2024 | |
|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Mean | *SD* |
| VR | *p* Value | 200 | 0.55 | 0.13 | 0.58 | 0.13 |
| | Point biserial | 200 | 0.29 | 0.05 | 0.30 | 0.06 |
| | IRT Difficulty | 200 | -0.21 | 0.66 | -0.20 | 0.67 |
| DM | Facility | 130 | 0.55 | 0.15 | 0.58 | 0.14 |
| | Point biserial | 130 | 0.36 | 0.11 | 0.38 | 0.11 |
| | IRT Difficulty | 130 | 0.23 | 0.69 | 0.21 | 0.69 |
| QR | *p* Value | 160 | 0.64 | 0.14 | 0.67 | 0.13 |
| | Point biserial | 160 | 0.41 | 0.07 | 0.45 | 0.07 |
| | IRT Difficulty | 160 | -0.29 | 0.74 | -0.27 | 0.73 |
| AR | *p* Value | 250 | 0.68 | 0.13 | 0.70 | 0.12 |
| | Point biserial | 250 | 0.34 | 0.07 | 0.37 | 0.07 |
| | IRT Difficulty | 250 | 0.16 | 0.69 | 0.15 | 0.67 |

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 56 summarises the number of items showing drift in the UCAT since 2017 and Table 57 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 2024, 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 56. Number of Operational Items Showing Drift in UCAT

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

Table 57. Number of Operational Items Showing Drift in UCAT ANZ

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

At present, it is recommended that the degree of drift is monitored in 2025. We would not recommend taking any action to create a separate item bank for the UCAT ANZ at this time.

## 7.2 SJT Item Analysis

Unlike the analysis undertaken on the cognitive sections, classical test statistics are sample-dependent, meaning that they are calculated based on the sample of candidates who respond to each item and are not linked back to a common benchmark group. Therefore, the item statistics presented for the SJT are not comparable to those presented for the cognitive sections due to the different measurement models used.

Prior to calculating the item statistics, outlier candidates are removed from the sample according to the criteria outlined in

Table 58. 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 58. Candidate Removal Summary for SJT Item Analysis

| Statistic | Criteria | Number of Candidates Removed |
|---|---|---|
| 1. *Z* score of the scaled score | *Z* score < -4.203 | 0 |
| 2. High number of missing responses | > 1 blank response on operational items | 1,159 |
| 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.
    - 0.1 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 were adjusted to align with those used for cognitive items. 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. This change supports the continued development and improvement of the item bank.

Table 59 shows the number of items that met and did not meet the quality criteria. The most/least item type was more successful than the standard items, with all operational items and 82% of the pretest items meeting the criteria.

Table 59. SJT Item Quality Criteria

| | Item Type | Statistical Criteria | All | | Appropriateness | | Direct Speech | | Importance | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | Rating Items | Met | 161 | 81% | 62 | 79% | 22 | 67% | 77 | 89% |
| | | Not met | 37 | 19% | 16 | 21% | 11 | 33% | 10 | 11% |
| | Most/Least Items | Met | 9 | 100% | | | | | | |
| | | Not met | 0 | 0% | | | | | | |
| Pretest | Rating Items | Met | 202 | 54% | 122 | 54% | 41 | 53% | 39 | 54% |
| | | Not met | 171 | 46% | 102 | 46% | 36 | 47% | 33 | 46% |
| | Most/Least Items | Met | 14 | 82% | | | | | | |
| | | Not met | 3 | 18% | | | | | | |

The proportion of items meeting the quality criteria has slightly improved compared to previous years, as shown in Figure 31. The number of pretest most/least items not meeting the criteria dropped from 37% in 2023 to 18% in 2024. In 2024, the percentage of standard rating items that did not meet the criteria was 46%, dropping from 57% from 2023 and 2022. It is likely that this increase in items meeting the quality criteria is due to the loosened criteria and potentially some slight improvement in item writing.

Figure 31. Proportion of SJT Items Failing Analysis 2017–2024

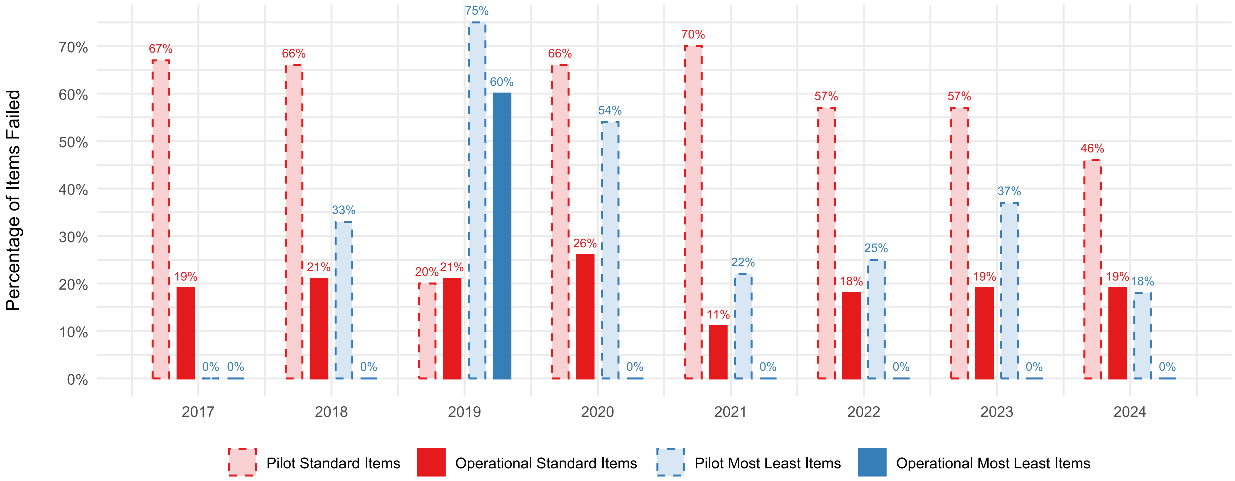


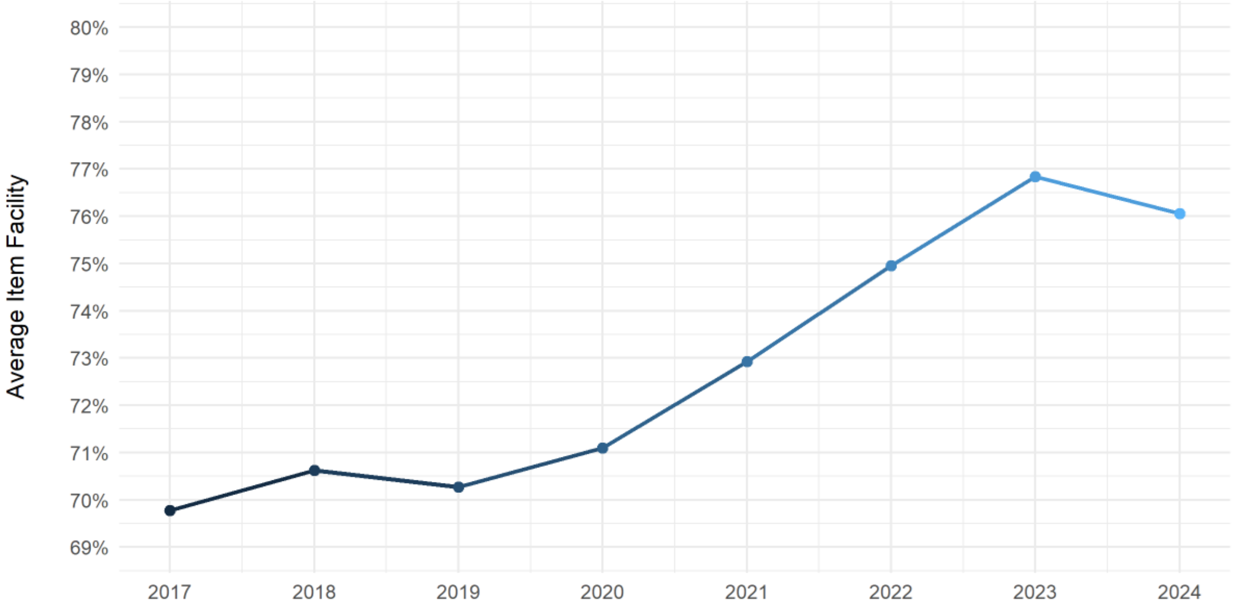
Table 60 provides a summary of the analysis of all operational SJT items.

Table 60. Operational SJT Item Analysis Summary

| | Mean | *SD* | Min | Max |
|---|---|---|---|---|
| Item Mean | 3.09 | 1.13 | 0.29 | 7.57 |
| Item *SD* | 1.03 | 0.31 | 0.33 | 2.20 |
| Item Partial Correlation | 0.26 | 0.12 | -0.02 | 0.53 |
| Item Total Facility | 0.76 | 0.17 | 0.10 | 0.98 |

Since 2017, the item mean score and facility have generally increased, as shown in Figure 32, indicating that items have become somewhat easier. Efforts have been made to increase item difficulty to balance this trend. It is encouraging to note that the SJT facility for this year is slightly lower than that observed last year.

Figure 32. Average Item Facility of Operational SJT Items 2017–2024



Unfortunately, Figure 33 shows a decrease in item partial correlation, indicating that, despite the test being slightly harder, its ability to discriminate between strong and weak candidates has declined. This suggests that the items were less effective overall at distinguishing candidate performance in comparison to 2022 and 2023. However, the item partial correlation remains well within the expected range.

Figure 33. Average Item Partial Correlation of Operational SJT Items 2017–2024

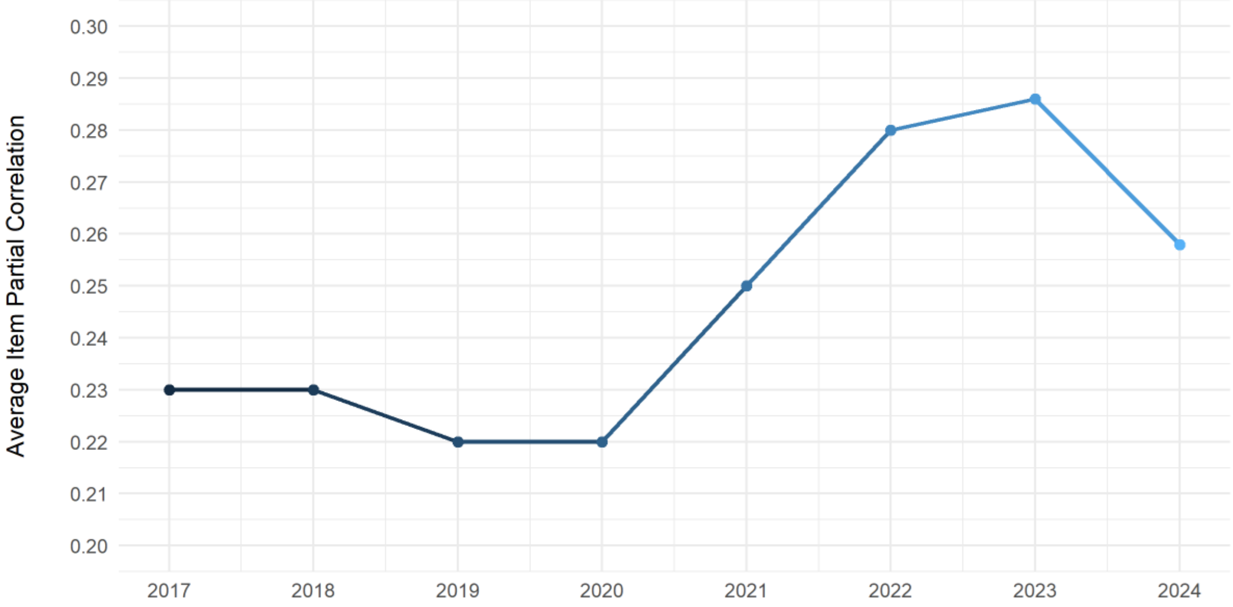


Table 61 summarises the statistics for the SJT pretest items. While the most/least items demonstrated slightly higher discriminating ability compared to the standard rating items, they also showed a relatively higher average item total facility.

Table 61. SJT Pretest Item Summary Statistics

|  | Statistic | Item Mean | Item *SD* | Item Partial | Item Total Facility |
|---|---|---|---|---|---|
| Rating Items | Mean | 2.76 | 0.91 | 0.14 | 0.74 |
|  | *SD* | 0.94 | 0.29 | 0.13 | 0.19 |
|  | Min | 0.65 | 0.09 | -0.17 | 0.22 |
|  | Max | 3.99 | 1.67 | 0.51 | 1.00 |
| Most/Least | Mean | 6.61 | 1.53 | 0.16 | 0.83 |
|  | *SD* | 1.37 | 0.54 | 0.09 | 0.17 |
|  | Min | 2.10 | 0.88 | -0.07 | 0.26 |
|  | Max | 7.60 | 2.46 | 0.27 | 0.95 |

## 7.3 Differential Item Functioning (DIF)

### 7.3.1 Introduction

DIF is a method for detecting potential bias in test items. For instance, if female and male candidates of the same ability level perform very differently on an item, then the item may be measuring something other than the ability of the candidates, possibly some characteristic of the candidates that is related to gender.

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

### 7.3.2 Method of DIF Detection

For the 2024 UCAT, a different method of DIF detection was employed for the cognitive sections and the SJT due to the different measurement models employed by the subtests. For the cognitive subtests, the Mantel-Haenszel procedure was used. This procedure compares the performance of different groups of candidates who are within the same ability strata. If there are overall differences between the groups for candidates of the same ability levels, then the item may be measuring something other than what it was designed to measure.

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

In both approaches, items were classified into one of three categories: A, B or C. Category A contains items with negligible DIF, Category B contains items with slight to moderate

DIF and Category C contains items with moderate to large DIF. For the cognitive subtests, these categories are derived from the DIF classification categories developed by Educational Testing Service (ETS) and are defined below:

A: DIF is not significantly different from zero or has an absolute value < 1.0
B: DIF is significantly different from zero and has an absolute value >= 1.0 and < 1.5
C: DIF is significantly larger than 1.0 and has an absolute value >= 1.5

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

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

### 7.3.3 Sample Size Requirements

Minimum sample-size requirements used for the UCAT DIF analyses were at least 50 candidate responses per group and at least 200 in total. If the sample size for the DIF analysis is less than 200, the sample is not large enough to undertake analysis and therefore DIF is not reported. Because pretest items were distributed across multiple versions of the forms, fewer responses are available per item than for operational items. As a result, it was not possible to compute DIF for many of the pretest items for certain group comparisons.

### 7.3.4 DIF Results

The DIF results are reported below for each demographic group. Table 62 shows DIF in relation to gender. One operational DM item was found to exhibit Category C DIF favouring Male over Female.

Table 62. Gender DIF

| Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 199 | 100% | 125 | 96% | 158 | 99% | 249 | 100% | 205 | 99% |
| | B | 1 | 0% | 4 | 3% | 2 | 1% | 1 | 0% | 2 | 1% |
| | C | 0 | 0% | 1 | 1% | 0 | 0% | 0 | 0% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |

| Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Pretest | A | 282 | 100% | 249 | 100% | 294 | 99% | N/A | N/A | 373 | 96% |
| | B | 1 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 17 | 4% |
| | C | 0 | 0% | 0 | 0% | 2 | 1% | N/A | N/A | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |

In 2024, the age comparison criteria were updated to increase the number of items where a comparison could be made. Since 2022, the comparison has been between candidates aged less than 20 and those aged greater than 25, as opposed to the previous comparison of less than 20 and greater than 35, as detailed in Table 63. One operational VR item was identified with Category C DIF, favouring older candidates. In DM, nine Category C DIF items were identified, with three favouring older candidates and six favouring younger candidates. Four QR items exhibited Category C DIF, with two favouring older candidates and two favouring younger candidates. For AR, two items showed Category C DIF, both favouring younger candidates.

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.

A further investigation was conducted to understand why the increase in Category C DIF items was not observed in 2022 but gradually rose in 2023 and 2024. It was found that the younger group (aged less than 20) is becoming increasingly international, with the proportion of UK candidates in this group decreasing from 82% in 2022 to 78% in 2023, and 75% in 2024. In contrast, the older group (aged more than 25) has maintained a relatively stable proportion of UK candidates, ranging between 89% and 90% across the years. The difference in DIF may be a reflection of the changing population, particularly the increase in younger international students. We will continue to monitor this trend. The items showing DIF will be reviewed by the Content Team and subsequently removed from the item bank.

Table 63. Age DIF

| Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 196 | 98% | 115 | 88% | 154 | 96% | 245 | 98% | 206 | 100% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | B | 3 | 2% | 6 | 5% | 2 | 1% | 3 | 1% | 1 | 0% |
| | C | 1 | 0% | 9 | 7% | 4 | 2% | 2 | 1% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | A | 0 | 0% | 3 | 1% | 0 | 0% | N/A | N/A | 380 | 97% |
| | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 9 | 2% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 1 | 0% |
| | NA | 283 | 100% | 246 | 99% | 296 | 100% | N/A | N/A | 0 | 0% |

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 64 identifies four instances of Category C DIF in the ethnicity comparisons for operational items. All four instances are linked to the same DM item, which favoured White candidates over Black, Asian, and Mixed candidates, and overall favoured White candidates over Non-White candidates.

Table 64. Ethnicity DIF

| Type | Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | White/ Black | A | 199 | 100% | 126 | 97% | 157 | 98% | 248 | 99% | 192 | 93% |
| | | B | 1 | 0% | 3 | 2% | 3 | 2% | 2 | 1% | 15 | 7% |
| | | C | 0 | 0% | 1 | 1% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | White/ Asian | A | 197 | 98% | 126 | 97% | 159 | 99% | 250 | 100% | 200 | 97% |
| | | B | 3 | 2% | 3 | 2% | 1 | 1% | 0 | 0% | 7 | 3% |
| | | C | 0 | 0% | 1 | 1% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | White/ Mixed | A | 200 | 100% | 129 | 99% | 160 | 100% | 249 | 100% | 207 | 100% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | 1 | 0% | 0 | 0% |
| | | C | 0 | 0% | 1 | 1% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | White/ Non-White | A | 198 | 99% | 127 | 98% | 160 | 100% | 250 | 100% | 204 | 99% |
| | | B | 2 | 1% | 2 | 2% | 0 | 0% | 0 | 0% | 3 | 1% |
| | | C | 0 | 0% | 1 | 1% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | White/ Black | A | 11 | 4% | 23 | 9% | 4 | 1% | N/A | N/A | 35 | 9% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 1 | 0% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |
| | | NA | 272 | 96% | 226 | 91% | 292 | 99% | N/A | N/A | 354 | 91% |
| | White/ Asian | A | 283 | 100% | 146 | 59% | 295 | 100% | N/A | N/A | 377 | 97% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 10 | 3% |
| | | C | 0 | 0% | 1 | 0% | 1 | 0% | N/A | N/A | 2 | 1% |
| | | NA | 0 | 0% | 102 | 41% | 0 | 0% | N/A | N/A | 1 | 0% |
| | White/ Mixed | A | 0 | 0% | 3 | 1% | 0 | 0% | N/A | N/A | 17 | 4% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |

| Type | Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| | | NA | 283 | 100% | 246 | 99% | 296 | 100% | N/A | N/A | 373 | 96% |
| | White/ Non- White | A | 282 | 100% | 246 | 99% | 296 | 100% | N/A | N/A | 378 | 97% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 10 | 3% |
| | | C | 1 | 0% | 1 | 0% | 0 | 0% | N/A | N/A | 2 | 1% |
| | | NA | 0 | 0% | 2 | 1% | 0 | 0% | N/A | N/A | 0 | 0% |

Since 2022, comparisons between SEC1 and non-SEC1 candidates have been included to enable more comprehensive analyses. Three operational items were identified with Category C DIF: one VR item favoured SEC1 over SEC4, one QR item favoured SEC1 over SEC2, and one DM item favoured SEC4 over SEC1.

Table 65. SEC DIF

| Type | Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | SEC 1/2 | A | 200 | 100% | 130 | 100% | 159 | 99% | 250 | 100% | 207 | 100% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | C | 0 | 0% | 0 | 0% | 1 | 1% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | SEC 1/3 | A | 200 | 100% | 130 | 100% | 160 | 100% | 250 | 100% | 207 | 100% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | SEC1/4 | A | 199 | 100% | 129 | 99% | 160 | 100% | 250 | 100% | 207 | 100% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | C | 1 | 0% | 1 | 1% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | SEC 1/5 | A | 199 | 100% | 130 | 100% | 160 | 100% | 250 | 100% | 207 | 100% |
| | | B | 1 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | SEC 1/(2-5) | A | 200 | 100% | 130 | 100% | 160 | 100% | 250 | 100% | 207 | 100% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | SEC 1/2 | A | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 232 | 59% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 6 | 2% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |
| | | NA | 283 | 100% | 249 | 100% | 296 | 100% | N/A | N/A | 152 | 39% |
| | SEC 1/3 | A | 48 | 17% | 38 | 15% | 32 | 11% | N/A | N/A | 356 | 91% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 6 | 2% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 1 | 0% |
| | | NA | 235 | 83% | 211 | 85% | 264 | 89% | N/A | N/A | 27 | 7% |
| | SEC 1/4 | A | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 254 | 65% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 5 | 1% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |
| | | NA | 283 | 100% | 249 | 100% | 296 | 100% | N/A | N/A | 131 | 34% |

| Type | Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| | SEC 1/5 | A | 1 | 0% | 16 | 6% | 0 | 0% | N/A | N/A | 304 | 78% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 4 | 1% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |
| | | NA | 282 | 100% | 233 | 94% | 296 | 100% | N/A | N/A | 82 | 21% |
| | SEC 1/(2-5) | A | 283 | 100% | 178 | 71% | 296 | 100% | N/A | N/A | 386 | 99% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 4 | 1% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |
| | | NA | 0 | 0% | 71 | 29% | 0 | 0% | N/A | N/A | 0 | 0% |

As shown in Table 66, two Category C DIF items were identified in the comparison between candidates with an honours degree or higher and those without. Both items were pretest items (one QR and one DM), and both favoured candidates with degree-level education over those without.

Table 66. Honours Degree DIF

| Type | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 200 | 100% | 127 | 98% | 160 | 100% | 250 | 100% | 204 | 99% |
| | B | 0 | 0% | 3 | 2% | 0 | 0% | 0 | 0% | 3 | 1% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | A | 283 | 100% | 188 | 76% | 295 | 100% | N/A | N/A | 379 | 97% |
| | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 11 | 3% |
| | C | 0 | 0% | 1 | 0% | 1 | 0% | N/A | N/A | 0 | 0% |
| | NA | 0 | 0% | 60 | 24% | 0 | 0% | N/A | N/A | 0 | 0% |

Table 67 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 67. English as First Language DIF

| Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 200 | 100% | 130 | 100% | 160 | 100% | 250 | 100% | 207 | 100% |
| | B | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | A | 283 | 100% | 249 | 100% | 296 | 100% | N/A | N/A | 375 | 96% |
| | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 15 | 4% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |

As shown in Table 68, 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 QR item was found to favour non-UK residents over UK residents.

Table 68. Residency DIF

| Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 200 | 100% | 128 | 98% | 158 | 99% | 247 | 99% | 200 | 97% |
| | B | 0 | 0% | 2 | 2% | 2 | 1% | 3 | 1% | 7 | 3% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | A | 283 | 100% | 249 | 100% | 295 | 100% | N/A | N/A | 367 | 94% |
| | B | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 23 | 6% |
| | C | 0 | 0% | 0 | 0% | 1 | 0% | N/A | N/A | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | N/A | N/A | 0 | 0% |

Only a very small number of candidates (34) took the online version of the UCAT (see Section 4.4), making comparison impossible.

In conclusion, 35 Category C DIF items were identified in 2024, comprising 21 operational items and 14 pretest items. This marks a significant increase compared to 24 items in 2023 (8 operational and 16 pretest items), 10 in 2022, and 13 in 2021. The increase is primarily in operational items and largely attributed to age comparisons, with 16 operational items flagged for age-related DIF. As most DIF items are operational, the increase likely reflects the detection of DIF in items that had previously passed analysis, rather than changes in the quality of item writing. The updated age comparison criteria introduced in 2022, which broadened the comparison from candidates aged 20 or younger versus 35 or older to 20 or younger versus 25 or older, may have contributed to the detection of DIF by allowing a larger portion of the candidate population to be included. Excluding the 16 age-related DIF items would reduce the total number of Category C DIF items to 19, aligning more closely with previous years. Although the criteria were updated in 2022, the particularly pronounced increase in 2024 might be partly explained by changes in candidate demographics. The younger group (aged less than 20) has become increasingly international, with the proportion of UK candidates decreasing from 82% in 2022 to 75% in 2024. In contrast, the older group (aged 25 or older) has maintained a stable proportion of UK candidates (89–90%). The observed DIF difference may be influenced by shifts in the candidate population particularly the growing number of younger international students.  To ensure fairness, all identified DIF items have been removed from the item bank and will not be used in future tests. Further efforts will focus on reviewing these items and refining item development processes to minimise potential bias moving forward.

# 8. Summary

The 2024 test specification shows no major deviations from the previous year, aside from changes to the VR item ratio, and additional restrictions on the Pause-the-clock feature. QR and AR were each scaled down by 10 points like in 2023, and VR was again scaled up by 20 points. However, the effect of this rescaling was smaller than anticipated, suggesting that other factors may have counteracted the intended impact. For VR, changes were made to the item type ratio, with an increase in multiple choice questions to improve test discriminability. However, this adjustment may have also made the subtest more speeded, potentially offsetting the upward rescaling effect. The additional restriction on the Pause-the-clock feature was found to be effective in ensuring its fair and appropriate use.

There have been changes in candidate composition this year. Most notably, 2024 saw the highest number of candidates on record, reflecting the continued growth of the exam, likely driven in part by the increase in international partner universities. Consequently, there was also an increase in non-UK candidates, who now represent the second largest group when categorised alongside UK ethnic groups, following UK-Asian candidates. This shift may also explain the rise in candidates who do not identify English as their first or primary language. Additionally, a gradual trend has emerged, with fewer candidates applying for medicine programmes via UCAS and an increasing number applying for dentistry. Aside from these changes, the composition of most other demographic groups remains largely consistent with previous years.

The changes in candidate composition could also help explain the relatively unusual SJT banding distribution observed this year. The SJT banding is skewed, with more candidates categorised in the lower bands than the target and fewer candidates in the higher bands. This indicates that overall performance on the SJT subtest was lower than expected. One contributing factor may be the relatively strong SJT performance observed last year, which led to higher banding cutoffs being applied this year. Another factor could be the changes in candidate composition. Non-UK candidates, who consistently perform the lowest on the SJT, represent a larger proportion of the cohort this year, which may have contributed to the overall decline. Additionally, the decrease in candidates applying for medicine, as well as those with English as their first or primary language, may have further impacted overall performance, as both groups typically perform better across all subtests, including the SJT.

In addition to changes in the candidate sample, which inevitably introduce variability in performance and contribute to the diminished rescaling effect, changes in the speededness of the test were also observed. Specifically, QR and AR have become less speeded, while VR has become more speeded. Efforts in the test construction process to reduce the speededness of subtests have been effective, particularly for QR and AR. This reduction in speededness likely enabled candidates to perform better, thereby offsetting the intended downward rescaling effect. Conversely, the changes in VR, which involved including more multiple-choice items to improve test discriminability, appear to have increased the subtest's speededness. This may have negatively impacted candidate performance, thereby counteracting the upward rescaling applied to VR.

In the item analysis for the cognitive subtests, both operational and pretest items showed an improvement in passing rates. A particularly notable change is that QR pretest items this year are not only more discriminating but also easier. This reflects the ongoing effort to create more easy items for the QR item bank, which currently contains a higher proportion of difficult items. For the SJT subtest, an improvement in item passing rates was also observed. However, this may be partially attributed to the adjusted, more lenient item passing criteria introduced this year. Additionally, a slight decline in item partial correlation and facility was noted for operational SJT items this year.

The DIF analysis revealed an increase in items categorised as Category C, which are items that showed significant bias toward certain candidate groups. Notably, this increase was observed in operational items but not in pretest items. This suggests that the increase is likely due to changes in the candidate sample rather than the quality of item writing, as all operational items had previously been tested without showing significant DIF. The greater diversity in this year’s sample, driven by an increase in international candidates, may have enabled better detection of biases in items. The absence of an increase in DIF for pretest items highlights that improvements in item writing have effectively limited bias in newer items, even with a more diverse candidate sample that flagged more DIF in operational items.

Candidates requiring special accommodations continue to represent a very small proportion of the overall candidate pool. The UCATSEN group remains the largest among those receiving accommodations, with a trend of the performance differences between UCAT and UCATSEN candidates widening slightly over time. Following the introduction of usage limits, the Pause-the-clock feature no longer shows signs of misuse, as observed last year. The usage pattern indicates diverse applications of the feature, suggesting it effectively meets the varied needs of candidates. However, the feature appears underutilised, with not all eligible candidates using it, and those who do use it often not using the full time allowed. This suggests that the current accommodations are more than sufficient. Therefore, no further adjustments to the feature are needed at this time.

Apart from these changes, the results of the 2024 UCAT administration were broadly consistent with those of previous years. Other than the mentioned differences, the demographic composition of test-takers remained largely unchanged, and the corresponding group performance differences also remained stable. In terms of test quality, the test forms were reliable, with appropriately low measurement error, and the forms were balanced, with average scores across forms being largely consistent.

## 8.1 Recommendations

UCAT has decided to remove the AR subtest starting in 2025. The primary reasons for this change are that AR has lower predictive validity compared to the other subtests and is highly coachable due to its nature. The time previously allocated to AR will be redistributed across the remaining subtests, with DM receiving a slight expansion to include additional items. Further details regarding the removal of AR can be found on the UCAT website (https://www.ucat.ac.uk/about-ucat/ucat-2025/).

In 2025, one additional minute will be allocated to the QR and VR subtests. It is recommended to scale both subtests down by 10 points to compensate for the potential score increase resulting from the additional time. The item type ratio for the VR subtest should remain unchanged, given the observed increase in speededness. The added restriction on the Pause-the-clock feature has been effective and should remain in place without any further modifications.

The approach to setting SJT banding cutoffs could be further investigated to identify ways to stabilise banding and reduce fluctuations caused by annual cutoff adjustments. However, as no conclusive solution has been identified, it is recommended that discussions continue to explore this further.

Given the major restructuring of the test with the removal of AR, it is recommended to minimise additional changes to avoid further instability in the test. Hence, no further changes are recommended.

# References

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.