Pearson

# University Clinical Aptitude Test (UCAT)

## Technical Report
## Testing Interval: 11 July 2022 to 29 September 2022

**Prepared by:**
Pearson VUE
27 February 2023

## 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 2022 from 11 July 2022 to 29 September 2022. This report covers 36,374 exams that were delivered during that period, which is a small decrease (2%) in 2021. The exam was delivered in two modes: online and test centre. Online test delivery accounted for only 0.2% of candidates, so it is not possible to reliably compare results between these two groups.

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

Each exam consists of five subtests. Mean scaled scores were stable for Verbal Reasoning (VR), Quantitative Reasoning (QR), Decision Making (DM) and Abstract Reasoning (AR), changing by less than ten scaled score points since 2021. All the Situational Judgement Test (SJT) bands are within 4% deviation from the target proportions. The proportion of candidates falling into the lowest SJT band decreased from 17% in 2021 to 14% in 2022, which is closer to the intended value of 10%.

The 2022 UCAT consisted of five test forms. Reliabilities for the forms were good across the board and corresponding standard errors of measurement (*SEM*s) were satisfactorily low and consistent with previous years.

The cognitive subtests were speeded to a certain extent. Most candidates used all the available time and the average time used was very close to the available time. The speededness has reduced fractionally across the cognitive subtests following small changes to the timing of QR and AR and the consideration of item time in the form build of all cognitive subtests in 2022. Speededness reduced in the SEN exams where candidates have more time available. The SJT remains the least speeded subtest.

Demographic trends in 2022 were consistent with those of previous years, with higher scores being associated with higher socio-economic classification (SEC), white ethnicity, speaking English as a first language, and being a UK resident. Males tended to perform better than females on the cognitive subtests, but females outperformed males on the SJT.

Individual item analysis showed satisfactory quality for the majority of operational items. Pretesting is intended to identify poor-quality items before they enter the operational scored test, and therefore the pretest items ranged more broadly in quality and on the whole performed less well. Four cognitive operational items and 19 cognitive pretest items failed the quality criteria and were removed from the item bank, whereas 35 SJT operational items and 127 SJT pretest items failed. Additionally, two pretest items were removed due to potentially exhibiting bias.

# 2. Introduction

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

This report covers the 2022 UCAT that was delivered from 11 July 2022 to 29 September 2022. As outlined in Section 2, the exam consisted of five subtests ranging from 26 to 66 items each. The design of the exam has remained the same as in recent years with a small change to the timing of two of the subtests. One minute (and five pretest items) was removed from AR and one minute was added to QR in order to reduce the speededness of QR.

Section 3 describes the exam results in terms of candidate volumes, scaled scores, and SJT bands. It also reports exam results in reference to candidates who qualified for a SEN version of the exam, whether candidates applied for medicine or dentistry, the mode of delivery, and candidate demographic characteristics.

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

# 3. Exam Design 2022

The 2022 UCAT consisted of five balanced test forms, each with five subtests. Each subtest includes scored and unscored items as shown in Table 1 below, in addition to changes made in to the 2022 test.

Table 1. UCAT Exam Design

| Subtest | Scored Items | Unscored Items | Total Number of Items | Changes |
|---|---|---|---|---|
| VR | 10 testlets of 4 items | 1 testlet of 4 items | 44 | None |
| QR | 8 testlets of 4 items | 1 testlet of 4 items | 36 | Added 1 minute |
| AR | 10 testlets of 5 items | None | 50 | Removed 5 pretest items<br>Removed 1 minute |
| DM | 1 testlet of 26 items | 3 items | 29 | None |
| SJT | 20 testlets of 1 to 5 items | 1 testlet of 5 items<br>1 testlet of 1 item | 66 | Reduced items from 69 to 66 |

Candidates were allowed 120 minutes to answer a total of 225 items from the five subtests. There were four groups of candidates with extra time allowances in 2022. The timing and scoring of the SEN exams are explored in detail in Section 3.2.

There have been changes to the subtests in 2022. The subtest times for AR and QR were changed from 2021 as one minute was removed from AR and moved to QR to address speededness concerns on QR – five pretest items were also removed from AR as a result. The item times were considered in the form build to address this additional time as well as a small adjustment to the scaling to counteract any drift because of this change. In addition to this change, the introduction of a new, lengthier, item type resulted in the reduction of the number of items on the SJT from 69 to 66.

Raw scores in each cognitive subtest were transformed to a scaled score ranging from 300 to 900. SJT scaled scores ranged from 300 to 801. 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 below.

Table 2. SJT Band Scaled Score Range and Description

| Bands | Scaled Score Range | Intended Band Proportions | Narrative |
|---|---|---|---|
| Band 1 | 658–900 | 22% | Those in Band 1 demonstrated an excellent level of performance, showing similar judgement in most cases to the panel of experts. |
| Band 2 | 597–657 | 38% | Those in Band 2 demonstrated a good, solid level of performance, showing appropriate judgement frequently, with many responses matching model answers. |
| Band 3 | 507–596 | 30% | Those in Band 3 demonstrated a modest level of performance, with appropriate judgement shown for some questions and substantial differences from ideal responses for others. |
| Band 4 | 300–506 | 10% | The performance of those in Band 4 was low, with judgement tending to differ substantially from ideal responses in many cases. |

The 2022 UCAT was delivered in two modes: the OnVUE mode, where a candidate can take the test in their own home with an online proctor, or the test centre mode, where candidates take the test in a specially designed test centre. Only 69 candidates took the online version of the test (see Section 3.4).

# 4. Examination Results

## 4.1 Overall Exam Results

This report covers examination results for 36,374 candidates who took the UCAT during the period 11 July 2022 to 29 September 2022. Candidate volumes have increased each year from 2017 to 2021 but showed a small decrease of 2% from 2021 to 2022, as illustrated in Figure 1 below.

Figure 1. Candidate Volumes since 2017

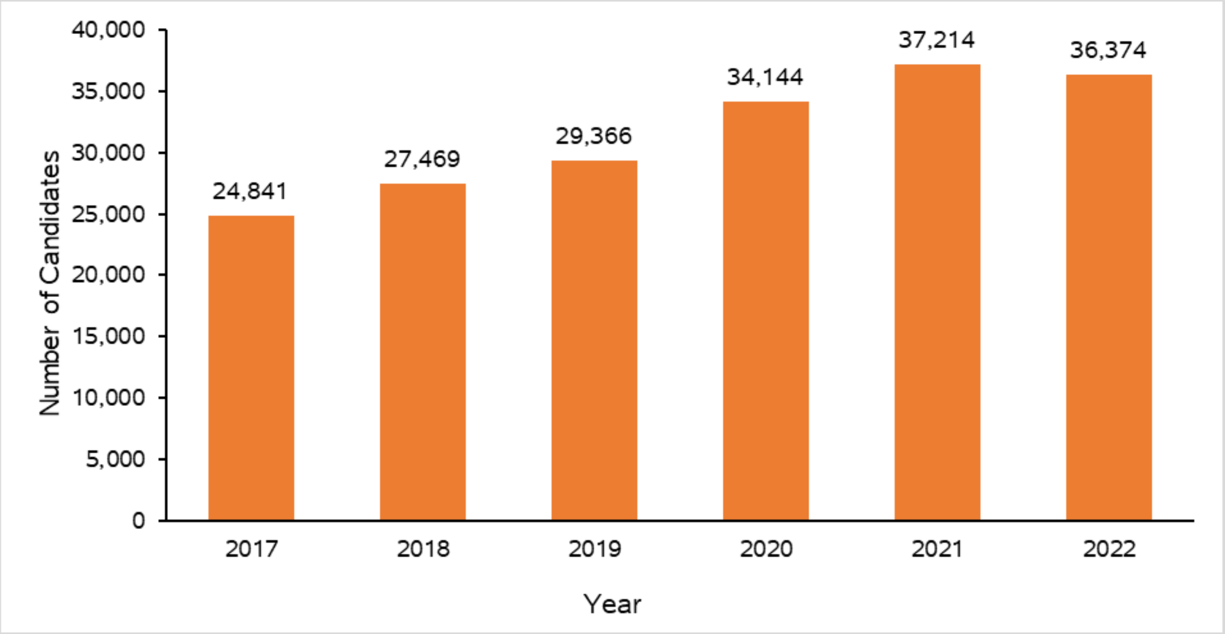


Table 3 presents summary statistics for each of the cognitive subtests plus the total scaled score for the cognitive subtests. VR scores were lowest with a mean score of 567, the highest average score was achieved on AR with an average of 659.

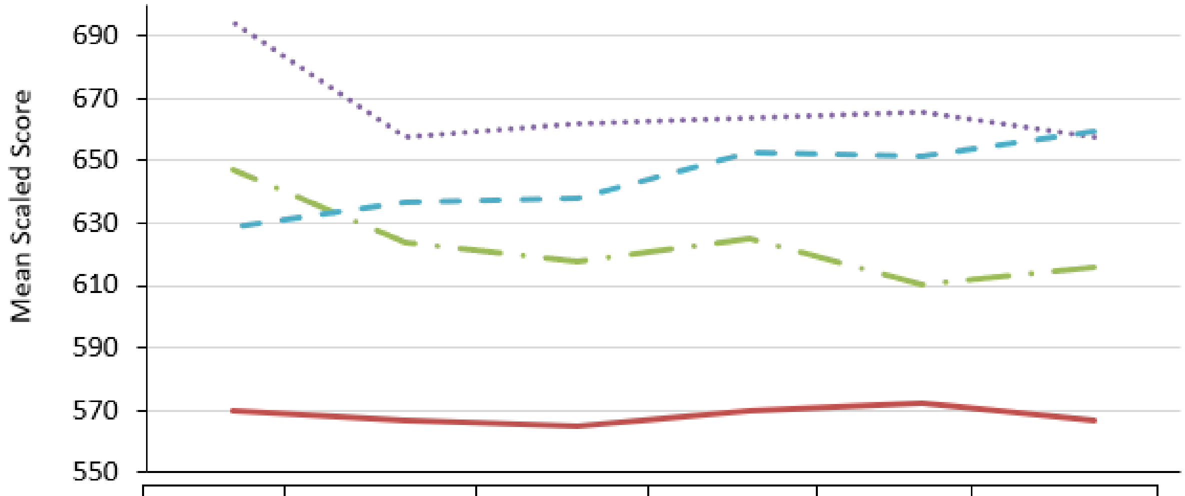
Table 3. Cognitive Subtest and Total Scaled Score Summary Statistics

| Subtest | Mean | *SD* | Min | Max |
|---|---|---|---|---|
| VR | 566.96 | 74.11 | 300 | 900 |
| DM | 615.88 | 91.52 | 300 | 900 |
| QR | 657.82 | 90.24 | 310 | 900 |
| AR | 659.41 | 98.72 | 300 | 900 |
| Total | 2,500.07 | 293.08 | 1,210 | 3,480 |

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

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

Figure 2. Scaled Scores by Year since 2017



| | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|---|
| VR | 570 | 567 | 565 | 570 | 572 | 567 |
| DM | 647 | 624 | 618 | 625 | 610 | 616 |
| QR | 695 | 658 | 662 | 664 | 665 | 658 |
| AR | 629 | 637 | 638 | 653 | 651 | 659 |

Since 2021, VR, DM, QR and AR have remained stable, changing by a small number of scaled score points which is well below one *SEM* for these subtests (as discussed in Section 5). This means that statistically the effect is not large enough to warrant concern. The change in the scaled scores for AR and QR is very small given that there have been changes to the timing of these subtests. When the test time is changed, it can make the form feel easier (with extra time) or harder (with less time), and this can lead to a shift in mean scores. However, this has not been the case, which means that the mitigations that were put in place were successful.

All of the subtests have showed 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 4.

Table 4. The Scaled Score Zero-Order Correlation of the Subtests

| | VR | DM | QR | AR |
|---|---|---|---|---|
| DM | 0.66*** | | | |
| QR | 0.57*** | 0.70*** | | |
| AR | 0.39*** | 0.54*** | 0.60*** | |
| SJT | 0.44*** | 0.53*** | 0.47*** | 0.48*** |

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

For the SJT, the number and percentage of candidates in each band for the 36,374 candidates who took the 2022 UCAT are shown in Table 5 below. Candidates are awarded a band for the SJT exam based on their underlying scaled score.

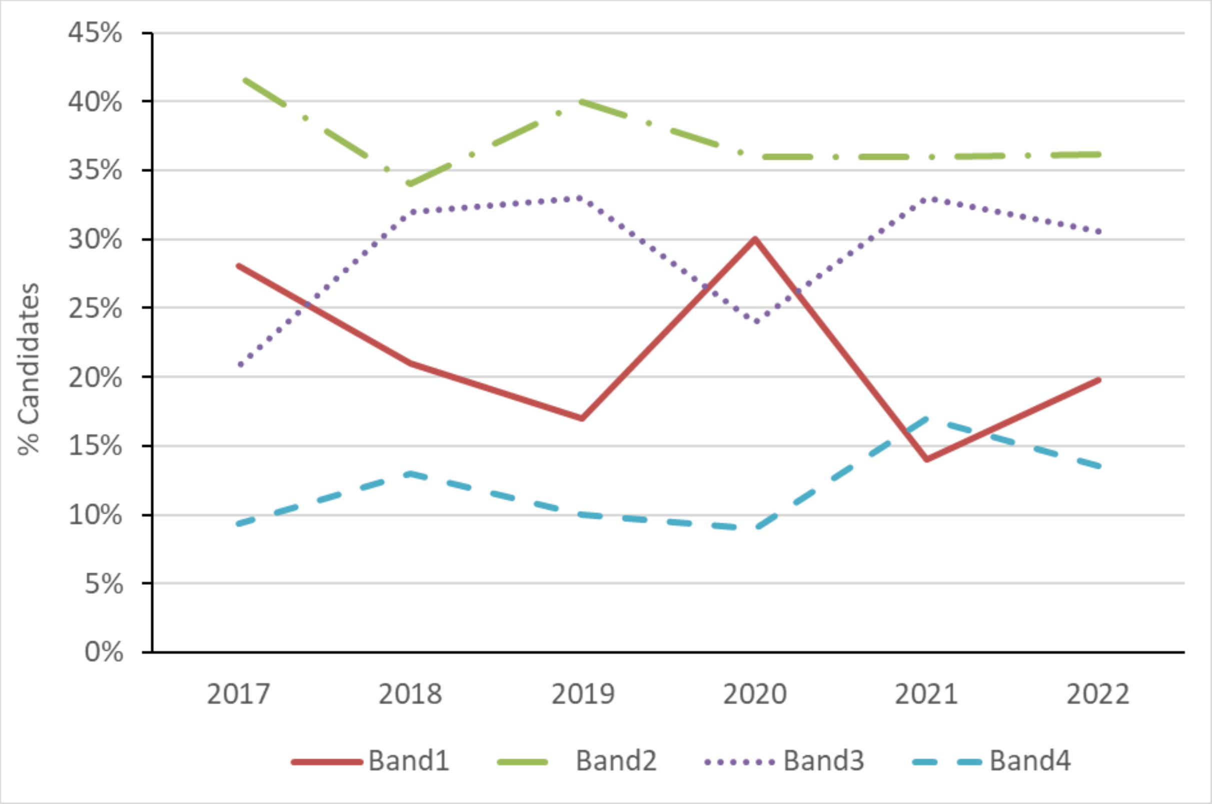
Table 5. SJT Band Distribution in 2022

| SJT Band | Number of Candidates | Mean Scaled Score | Percentage of Candidates | Target % |
|---|---|---|---|---|
| Band 1 | 7,184 | 681.07 | 19.75% | 22% |
| Band 2 | 13,142 | 627.50 | 36.13% | 38% |
| Band 3 | 11,128 | 559.08 | 30.59% | 30% |
| Band 4 | 4,920 | 439.88 | 13.53% | 10% |
| Total | 36,374 | 591.77 | 100.00% | 100% |

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

Each year the target proportion can change. Figure 3 illustrates the distribution of candidates across SJT bands since 2017.

Figure 3. SJT Band Proportions 2017–2022



The equating method undertaken when constructing test forms ensures that the difficulty of the test forms is controlled year-on-year, meaning test construction is not the source of the shifts in performance we see in Figure 3.

The distribution of scores is important because the band boundaries (defined in Table 2) are set each year by candidate performance in the prior year. Candidate performance in 2020 was relatively high, with an increase in candidates being categorised as Band 1. This increase resulted in the boundary for Band 1 being higher in 2021 than in 2020; therefore, when candidate performance returned to normal, correspondingly fewer candidates were categorised as Band 1. However, the 2022 band thresholds were based on the 2021 population, and therefore the band distributions are much closer to the target.

## 4.2 Special Educational Needs

There are four exams available for SEN candidates who are allowed extra time and breaks. Time allowances for each subtest and exam are illustrated below in Table 6.

Table 6. Exam Version Time Allowed

| Subtest | UCAT | UCATSEN | UCATSENSA | UCATSEN50 | UCATSA |
|---|---|---|---|---|---|
| VR | 00:21:00 | 00:26:15 | 00:26:15 | 00:31:30 | 00:21:00 |
| DM | 00:31:00 | 00:38:45 | 00:38:45 | 00:46:30 | 00:31:00 |
| QR | 00:25:00 | 00:31:15 | 00:31:15 | 00:37:30 | 00:25:00 |
| AR | 00:12:00 | 00:15:00 | 00:15:00 | 00:18:00 | 00:12:00 |
| SJT | 00:26:00 | 00:32:30 | 00:32:30 | 00:39:00 | 00:26:00 |

Only 6% of candidates took a SEN version of the exam, which is consistent with 2021. The most popular SEN exam was UCATSEN, as shown in Table 7 below. These exams are available to candidates who require additional time due to a special accommodation.

Table 7. Exam Version Candidate Volumes

| Exam | *N* | % |
|---|---|---|
| UCAT | 34,181 | 94% |
| UCATSEN | 1,605 | 5% |
| UCATSENSA | 329 | 1% |
| UCATSEN50 | 131 | 0% |
| UCATSA | 128 | 0% |
| Total | 36,374 | 100% |

Historically, candidates who take a SEN version of the exam usually outperform candidates who take the non-SEN version. Table 8 summarises the scaled score statistics by exam version. SEN candidates outperformed non-SEN candidates in all four subtests. The sample size of UCATSEN50, UCATSA, and UCATSENSA are small and results for those versions should be treated with caution.

Table 8. SEN and Non-SEN Cognitive Subtest

| Subtest | Statistic | UCAT (34,181) | UCATSEN (1,605) | UCATSENSA (329) | UCATSEN50 (131) | UCATSA (128) |
|---|---|---|---|---|---|---|
| VR | Mean | 565.40 | 585.78 | 611.19 | 610.69 | 588.75 |
| | *SD* | 73.34 | 80.77 | 79.25 | 93.92 | 72.69 |
| | Min | 300 | 350 | 420 | 430 | 400 |
| | Max | 900 | 900 | 900 | 900 | 890 |
| DM | Mean | 614.42 | 634.67 | 651.82 | 649.77 | 643.83 |
| | *SD* | 91.24 | 93.99 | 89.98 | 85.07 | 91.44 |
| | Min | 300 | 320 | 400 | 410 | 460 |
| | Max | 900 | 890 | 890 | 880 | 870 |
| QR | Mean | 656.12 | 680.12 | 701.58 | 691.83 | 683.91 |
| | *SD* | 89.86 | 91.19 | 96.17 | 96.82 | 83.62 |
| | Min | 310 | 370 | 410 | 480 | 500 |
| | Max | 900 | 900 | 900 | 900 | 880 |
| AR | Mean | 657.68 | 683.83 | 695.23 | 699.62 | 682.42 |
| | *SD* | 98.69 | 93.12 | 101.21 | 107.26 | 93.06 |
| | Min | 300 | 300 | 300 | 400 | 480 |
| | Max | 900 | 900 | 890 | 890 | 880 |
| Total | Mean | 2,493.62 | 2,584.40 | 2,659.82 | 2,651.91 | 2,598.91 |
| | *SD* | 291.77 | 291.47 | 302.30 | 315.34 | 279.67 |
| | Min | 1,210 | 1,540 | 1,760 | 1,830 | 2,010 |
| | Max | 3,480 | 3,420 | 3,470 | 3,430 | 3,270 |

Table 8 also includes the mean total cognitive scaled score for each exam version. It is evident that SEN candidates performed better than non-SEN candidates on the cognitive subtests as a whole. The difference between candidates who sat the UCAT and those who sat the UCATSEN amounts to 91 scaled score points, this is higher than in 2021 where the difference was 67 scaled score points.

The pattern of SEN candidates being stronger than non-SEN candidates is repeated for the SJT results, where the UCAT version of the exam has the lowest proportion of candidates in Band 1 and the highest in Band 4. The breakdown of SJT band proportions by exam version is presented in Table 9 below. The version of the exam on which candidates performed the best is the UCATSEN50, where 31% of candidates are categorised as Band 1 and 8% are categorised as Band 4, but note the prior warning that few candidates sat that version of the exam, meaning comparison may not be reliable.

Table 9. SJT Band by Exam Version

| Exam Version | Mean Scaled Score | Band 1 | Band 2 | Band 3 | Band 4 |
|---|---|---|---|---|---|
| UCAT | 590.76 | 19.47% | 35.94% | 30.72% | 13.87% |

| | | | | | |
|---|---|---|---|---|---|
| UCATSEN | 605.06 | 23.24% | 38.63% | 29.41% | 8.72% |
| UCATSENSA | 614.89 | 25.53% | 41.34% | 27.66% | 5.47% |
| UCATSEN50 | 616.02 | 30.53% | 37.40% | 23.66% | 8.40% |
| UCATSA | 612.66 | 25.00% | 40.63% | 25.78% | 8.59% |

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

Table 10. Stratified Sample of 2020 UCAT

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

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

## 4.3 Medicine and Dentistry

Many candidates who take the UCAT also apply for medical or dental school via the Universities and Colleges Admissions Service (UCAS). This section of the report concerns the performance of candidates in relation to whether they apply 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 63% of candidates, a reduction from 69% in 2021. Eleven percent of candidates applied for dentistry (increased from 9% in 2021), and the remaining 26% applied for neither or could not be matched with UCAS data.

Candidates who applied for medicine as a first choice outperformed those who applied for dentistry, as illustrated in Table 11. The highest mean scaled score was achieved on AR and the lowest on VR for both candidate groups. Candidates who did not apply for medicine or dentistry or were not matched by UCAS performed less well than both other groups.

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

| | Mean | | | *SD* | | |
|---|---|---|---|---|---|---|
| Subtest | Medicine | Dentistry | None | Medicine | Dentistry | None |
| VR | 584.13 | 559.64 | 529.02 | 72.33 | 65.82 | 66.46 |
| DM | 640.39 | 614.11 | 558.26 | 85.87 | 82.66 | 81.43 |
| QR | 681.17 | 659.59 | 601.53 | 87.85 | 81.75 | 72.64 |
| AR | 682.61 | 668.80 | 600.43 | 95.40 | 92.28 | 83.54 |
| Total | 2,588.30 | 2,502.14 | 2,289.23 | 273.13 | 257.61 | 240.12 |

Better performance by medicine candidates is also reflected in the SJT banding. As Table 12 shows, more medicine than dentistry candidates appeared in Band 1, and fewer medicine than dentistry candidates appeared in Band 4.

Table 12. Medicine/Dentistry Candidates: SJT Bands

| Group | Mean Scaled Score | Band 1 | Band 2 | Band 3 | Band 4 |
|---|---|---|---|---|---|
| Dentistry | 602.41 | 21% | 39% | 32% | 8% |
| Medicine | 610.70 | 24% | 41% | 28% | 6% |
| None | 542.45 | 8% | 23% | 37% | 32% |

In summary, UCAT candidates who apply for medicine perform 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 2022, the UCAT was offered in both the standard test centre and online proctored mode. Only 69 candidates took the exam in the online proctored mode, amounting to only 0.2% of all candidates. This contrasts with 2020, when more than 11,038 candidates took the exam in the online mode, and 231 in 2021. 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 2022, it is not possible to draw reliable inferences on differences in performance for the 2022 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 analysed the performance of individual items by demographic characteristic and removed any items that might exhibit bias (as discussed in Section 6.3).

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 cognitive subtests scaled scores.

### 4.5.2 Gender

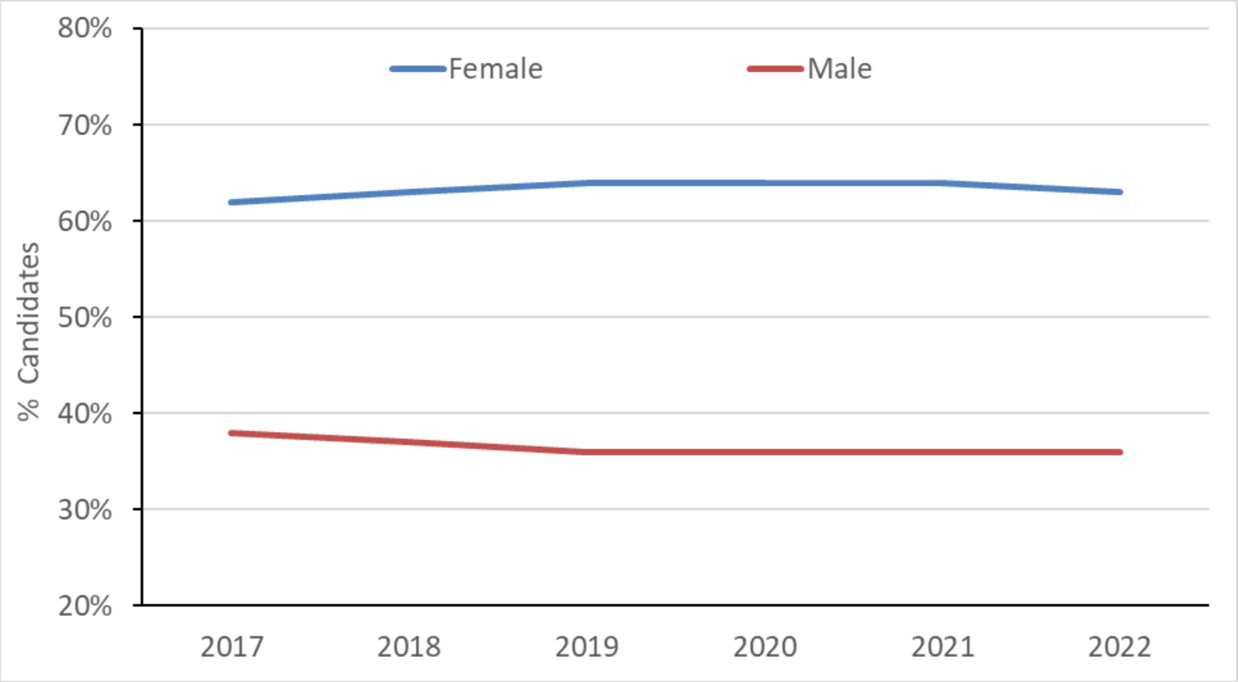
Table 13 presents the breakdown of test-takers by gender. The majority of test-takers were female, and only 236 stated “Other”, or that they would prefer not to say.

Table 13. Gender Counts

| Gender | *N* | % |
|---|---|---|
| Female | 22,908 | 63% |
| Male | 13,230 | 36% |
| I prefer not to say | 172 | 1% |
| Other | 64 | 0% |

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

Figure 4. Distribution of Candidates by Gender 2017–2022



Males outperformed females on all subtests except the SJT, where females performed better than males. The difference between male and female average scores is shown in Table 14, ranging from 9 scaled score points on VR to 33 scaled score points on QR. However, note that these differences are less than the *SEM* on the subtest and therefore may not be significant. Further analysis can be found below.

Table 14. Gender Scaled Scores

| Subtest | Mean Scaled Score | | *SD* Scaled Score | |
|---|---|---|---|---|
| | Female | Male | Female | Male |
| VR | 563.36 | 572.42 | 73.54 | 74.24 |
| DM | 608.81 | 627.53 | 90.82 | 91.60 |
| QR | 645.47 | 678.96 | 86.61 | 92.58 |
| AR | 655.62 | 665.83 | 96.99 | 101.44 |
| Total Cognitive | 2,473.25 | 2,544.74 | 288.40 | 295.72 |
| SJT | 597.27 | 582.16 | 76.95 | 82.36 |

A statistical test was used to examine whether the differences between the two groups observed in Table 14 were statistically significant. Table 15 shows the *T*-statistic, degrees of freedom and *p* value for each subtest and the total cognitive scores. The *df* column shows the available sample size. A non-zero *T*-statistic indicates there is a difference in the mean scaled score between two group samples. However, the difference may or may not be statistically significant. That is, the difference may or may not be sufficient evidence of a true difference in the entire population (e.g., between all eligible males and all eligible females). The *p* value shows the probability due to chance of observing a particular

*T*-statistic (or something more extreme). Lower *p* values (e.g., less than 0.01) indicate that we would be unlikely to see such a difference in our sample if there were no true difference in the population.

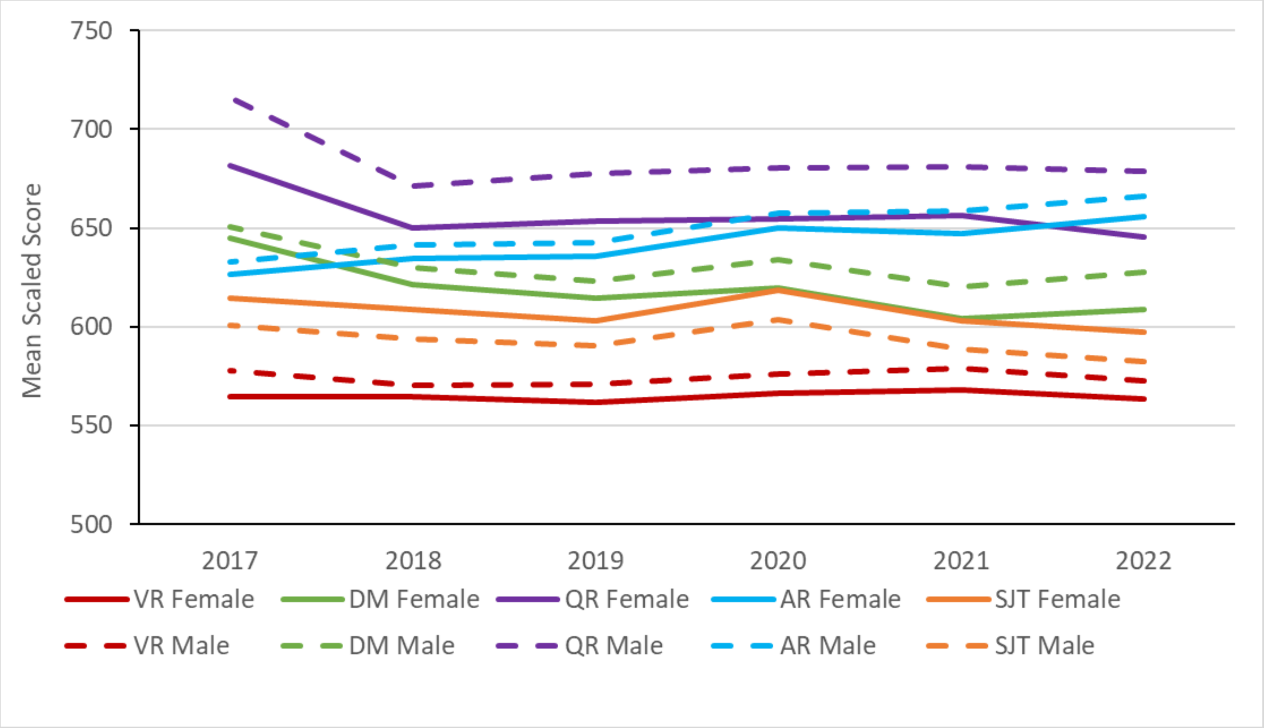
Therefore, Table 15 shows us that there are differences between male and female performance on each subtest and on the total cognitive scores, and that these differences are likely not to be the result of random chance.

Table 15. Gender *T*-Test

| Subtest | *T*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 9.06 | 36,136 | < 0.01 |
| DM | 18.73 | 36,136 | < 0.01 |
| QR | 33.49 | 36,136 | < 0.01 |
| AR | 10.22 | 36,136 | < 0.01 |
| Total Cognitive | 71.49 | 36,136 | < 0.01 |
| SJT | -15.11 | 36,136 | < 0.01 |

Figure 5 illustrates the subtest differences by gender. Differences have tended to be consistent year on year. Since 2017, the difference in scores between males and females has slightly broadened in the DM subtest. Since 2020, the range has slightly broadened in the AR subtest, although this has narrowed again in 2022. Interestingly, the gap for QR has increased in 2022 compared to 2021.

Figure 5. Scaled Score Distribution of Candidates by Gender 2017–2022



### 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 have been 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 is no longer a separate category. The categories used are:

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

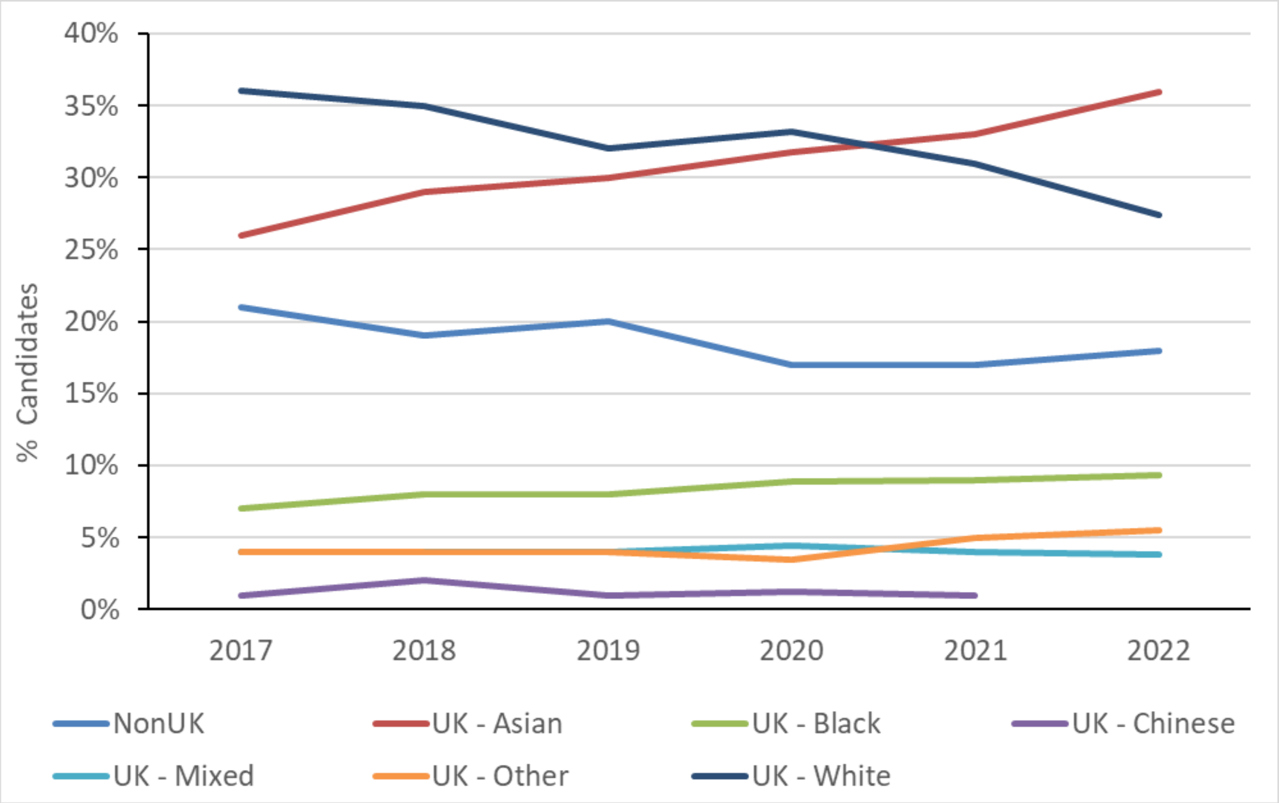
Table 16 shows the breakdown of candidates by ethnicity in the 2022 exam. The biggest candidate group was UK – Asian. Seventeen percent of candidates were not categorised due to being non-UK candidates.

Table 16. Ethnic Group Counts

| Country | Ethnic Group | *N* | Percent UK Candidates | Percent Total Candidates |
|---|---|---|---|---|
| UK | Asian | 12,748 | 44% | 36% |
| UK | White | 9,701 | 33% | 27% |
| UK | Black | 3,302 | 11% | 9% |
| UK | Other ethnic group | 1,952 | 7% | 6% |
| UK | Mixed | 1,356 | 5% | 4% |
| Non-UK | Non-UK | 6,370 | NA | 18% |

The proportion of candidates in each ethnic group has remained fairly stable in recent years. Figure 6 shows that the most common ethnic group changed from White to Asian for the first time in 2021 Since then, the proportion of White candidates has continued to decrease, and “UK – Asian” has remained the most common ethnic group. The proportion of non-UK candidates has decreased since 2017 and the proportion of Black candidates has slightly increased. Note that in 2022, “UK – Chinese” was not an option on the survey.

Figure 6. Distribution of Candidates by Ethnic Group 2017–2022



UK – White candidates performed better on average on all subtests than other groups. Table 17 shows the average scores in each subtest for each ethnic group. Performance was the lowest for UK – Black candidates on average on all subtests except the SJT, where non-UK candidates received the lowest average scaled scores.

Table 17. Ethnic Group Mean Scaled Score

| Subtest | Asian | White | Black | Other Ethnic Group | Mixed | Non-UK |
|---|---|---|---|---|---|---|
| VR | 561.59 | 592.42 | 548.06 | 544.31 | 583.41 | 552.23 |
| DM | 610.78 | 648.06 | 583.39 | 591.09 | 631.63 | 598.64 |
| QR | 663.86 | 676.25 | 619.39 | 643.08 | 665.13 | 640.53 |
| AR | 667.05 | 675.31 | 627.35 | 654.51 | 667.50 | 636.09 |
| Total Cognitive | 2,503.27 | 2,592.03 | 2,378.20 | 2,433.00 | 2,547.66 | 2,427.49 |
| SJT | 596.92 | 609.96 | 585.71 | 585.34 | 604.58 | 554.95 |

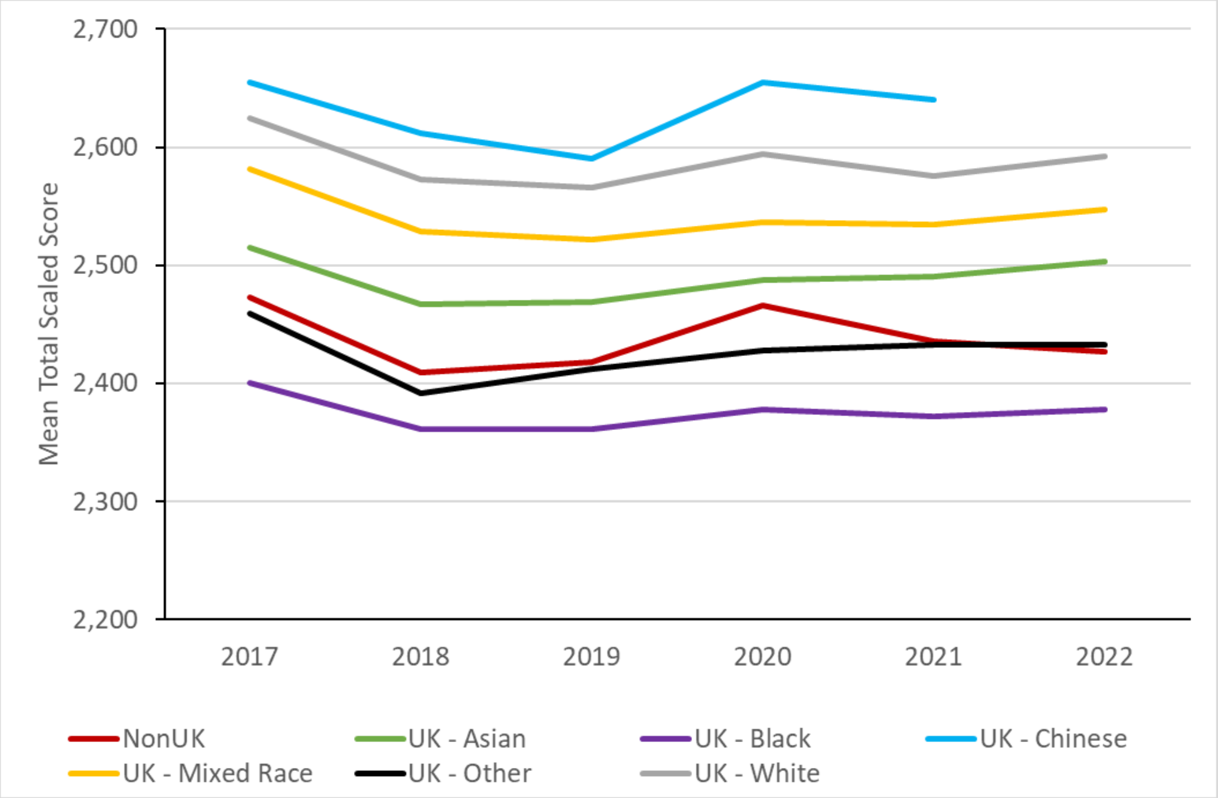
An *F*-test was used to examine whether the differences observed in Table 16 were likely to be due to chance. An *F*-test is similar to the *T*-test discussed in relation to gender (see 3.5.2). It is used when there are more than two groups. Table 18 has a positive *F*-statistic for each subtest and a *p* value of less than 0.01, which indicates that the differences observed in Table 17 above are likely to reflect true differences in performance in the candidate population.

Table 18. Ethnic Group *F*-Test

| Subtest | *F*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 339.10 | 6 | < 0.01 |
| DM | 364.92 | 6 | < 0.01 |
| QR | 234.56 | 6 | < 0.01 |
| AR | 179.41 | 6 | < 0.01 |
| Total Cognitive | 363.20 | 6 | < 0.01 |
| SJT | 355.66 | 6 | < 0.01 |

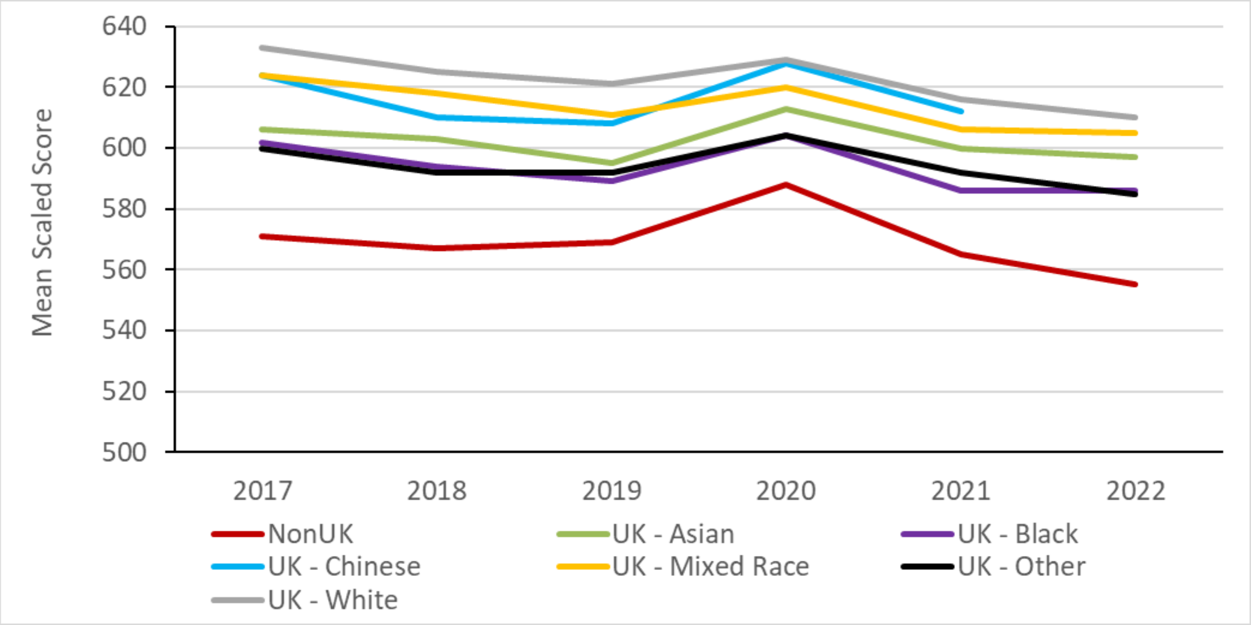
Mean total cognitive scaled scores fell for all ethnic groups between 2017 and 2018 reflecting the rescaling that took place (Figure 7). After 2018, scores remained fairly stable for UK White, UK Mixed Race and UK Black ethnic groups, with small increases for UK Asian and UK Other. The UK-Chinese ethnic category was removed from the survey in 2022.

Figure 7. Ethnic Group Mean Scaled Score for Total Scaled Score 2017–2022



In the SJT, there was a fairly large increase in scores for all ethnic groups between 2019 and 2020 and a slightly larger fall for all groups between 2020 and 2022. The most notable thing about ethnic group trends for the SJT is the margin by which non-UK candidates underperform relative to the other groups, as can be observed in Figure 8.

Figure 8. Ethnic Group Mean Scaled Score for SJT 2017–2022



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

### 4.5.4 Socio-Economic Classification

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

This issue is illustrated in Table 19, which shows that 20% of all candidates reside in the UK but cannot be categorised into an SEC. The candidates who can be categorised fall predominantly into SEC 1, representing Managerial and Professional Occupations.

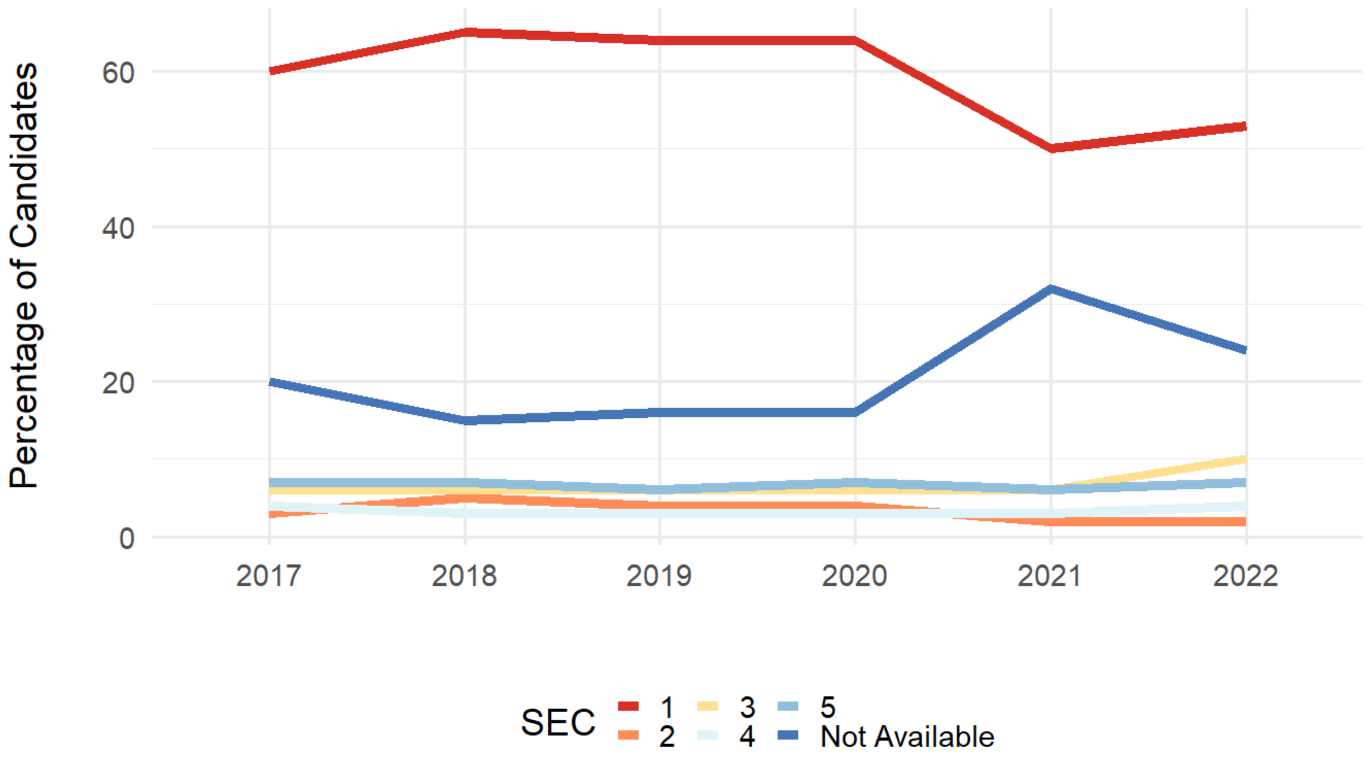
Table 19. SEC Counts

| Country | SEC | *N* | % of SEC | % of All |
|---|---|---|---|---|
| UK | 1 | 15,871 | 53% | 44% |
| | 2 | 584 | 2% | 2% |
| | 3 | 3,071 | 10% | 8% |
| | 4 | 1,141 | 4% | 3% |
| | 5 | 2,068 | 7% | 6% |
| | Unknown | 7,269 | 24% | 20% |
| EU | | 1,056 | | 3% |
| Other | | 5,314 | | 15% |

*Note.* Codes for NS-SEC Groups
1 – Managerial and Professional Occupations
2 – Intermediate Occupations
3 – Small Employers and Own Account Workers
4 – Lower Supervisory and Technical Occupations
5 – Semi-routine and Routine Occupations
NA – Could not calculate SEC group, i.e. information withheld

Prior to 2021, SEC was calculated for up to two parents or carers, then candidates were categorised as the highest of the two SECs. However, in 2021 the SEC questions changed to ask candidates to enter responses for only the highest earning parent or carer. The result is that proportionally more candidates appear in the NA category from 2021 than in previous years, as illustrated in Figure 9. It suggests that there are fewer candidates in SEC 1 since 2021 than in previous years; however, since this fall corresponds to a similar rise in SEC NA, it is likely that the new way of measuring SEC is influencing this measure. The trend in 2022 is similar to those observed in 2021.

Figure 9. Candidates by SEC 2017–2022



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 20.

Table 20. SEC Scaled Scores

| Mean Scaled Score | | | | | | |
|---|---|---|---|---|---|---|
| Subtest | SEC 1 | SEC 2 | SEC 3 | SEC 4 | SEC 5 | NA |
| VR | 581.79 | 572.45 | 562.50 | 562.80 | 552.53 | 553.69 |
| DM | 635.80 | 611.70 | 607.00 | 608.07 | 595.45 | 598.62 |
| QR | 674.75 | 646.76 | 651.96 | 648.89 | 641.41 | 645.41 |
| AR | 676.37 | 641.68 | 656.36 | 648.36 | 643.86 | 651.70 |
| Total Cognitive | 2568.72 | 2472.59 | 2477.82 | 2468.12 | 2433.25 | 2449.42 |
| SJT | 608.18 | 602.61 | 595.01 | 589.84 | 590.05 | 586.79 |
| *SD* | | | | | | |
| VR | 73.33 | 73.75 | 69.81 | 68.09 | 66.73 | 71.26 |
| DM | 88.31 | 86.15 | 86.30 | 86.29 | 85.05 | 90.50 |
| QR | 88.52 | 83.68 | 80.87 | 84.55 | 85.71 | 88.19 |
| AR | 96.53 | 95.58 | 93.35 | 95.37 | 91.84 | 97.11 |
| Total Cognitive | 282.45 | 273.10 | 266.81 | 270.24 | 266.10 | 288.00 |
| SJT | 68.39 | 69.35 | 72.58 | 74.53 | 76.50 | 81.59 |

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

Table 21. SEC *F*-Test

| Subtest | *F*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 193.88 | 5 | < 0.01 |
| DM | 237.18 | 5 | < 0.01 |
| QR | 160.06 | 5 | < 0.01 |
| AR | 111.03 | 5 | < 0.01 |
| Total Cognitive | 250.78 | 5 | < 0.01 |
| SJT | 102.26 | 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 22). 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 and 81% in 2022.

Table 22. Age Counts

| Age | *N* | Percent |
|---|---|---|
| <= 15 | 60 | 0% |
| 16–19 | 29,475 | 81% |
| 20–24 | 5,155 | 14% |
| 25–34 | 1,451 | 4% |
| >= 35 | 224 | 1% |

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

Figure 10. Average Scaled Scores by Age

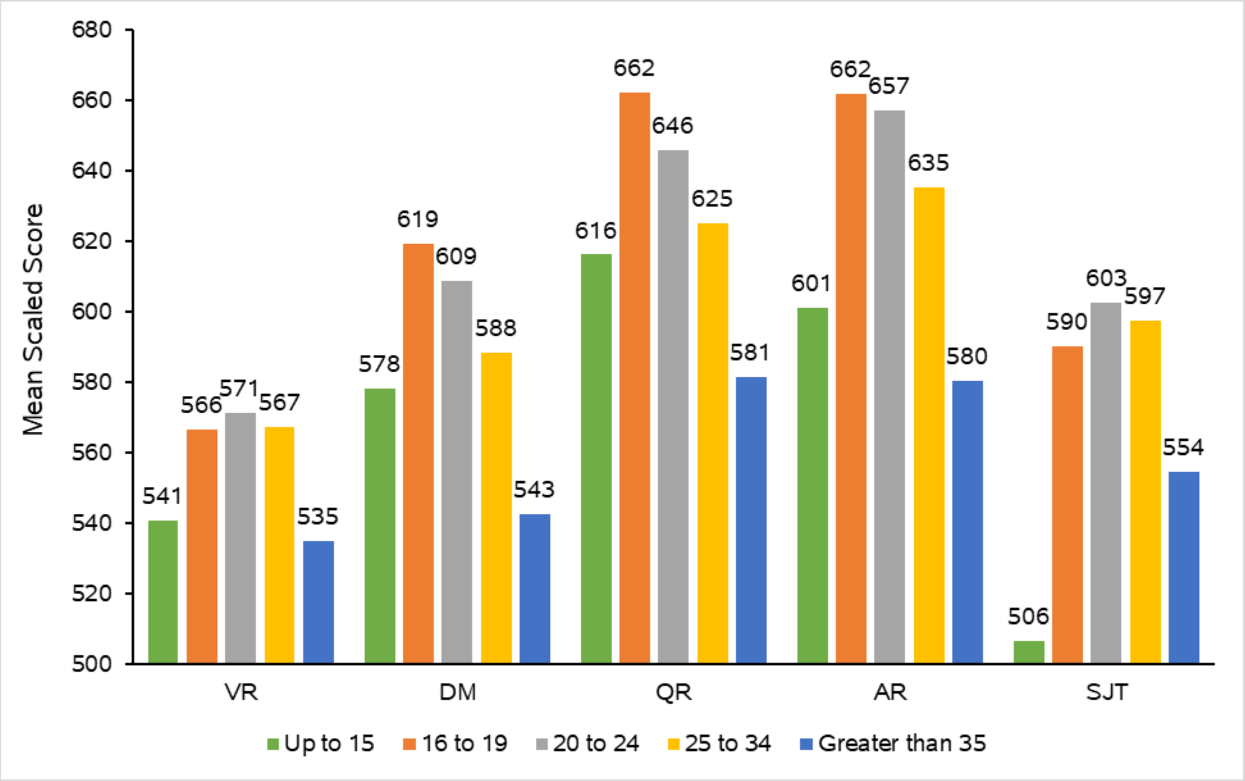
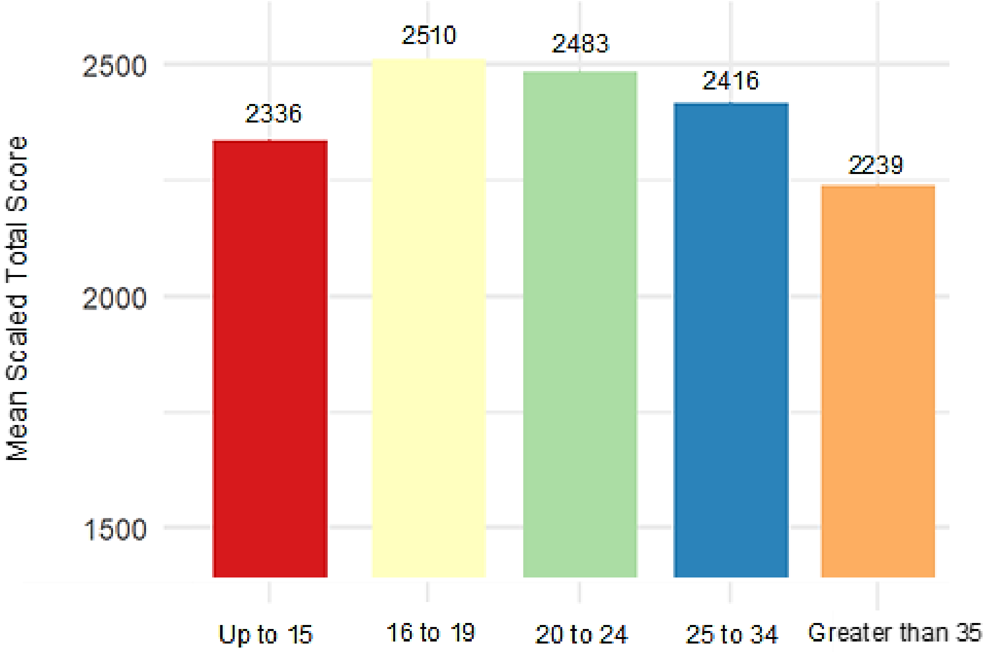


Figure 11. Average Scaled Total Scores of the Cognitive Subtests by Age



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

Table 23. Age *F*-Test

| Subtest | *F*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 16.95 | 4 | < 0.01 |
| DM | 89.79 | 4 | < 0.01 |
| QR | 134.01 | 4 | < 0.01 |
| AR | 68.76 | 4 | < 0.01 |
| Total | 92.54 | 4 | < 0.01 |
| SJT | 59.76 | 4 | < 0.01 |

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

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

| Subtest | Correlation | Significance |
|---|---|---|
| VR | -0.0112 | *p* > 0.01 |
| DM | -0.0972 | *p* < 0.01 |
| QR | -0.1153 | *p* < 0.01 |
| AR | -0.0727 | *p* < 0.01 |
| Total Cognitive | -0.0931 | *p* < 0.01 |
| SJT | 0.0179 | *p* < 0.01 |

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

### 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 2022 had a school leaver qualification (82%), 16% had a degree or above (down from 19% in 2021), and a small minority had no formal qualifications.

Candidates with a degree or above performed better on average on VR and the SJT. For the other cognitive subtests and the total cognitive score, below-honours degree candidates performed better on average, as shown in Table 25.

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

Table 25. Education Scaled Scores

| Subtest | School Leaver Qualification | Degree Level or Above |
|---|---|---|
| | Mean Scaled Score | |
| *N* | 29,899 | 5,826 |
| VR | 566.53 | 572.39 |
| DM | 618.86 | 605.39 |
| QR | 661.91 | 641.18 |
| AR | 661.86 | 651.57 |
| Total Cognitive | 2,509.06 | 2,470.53 |
| SJT | 590.65 | 602.51 |
| | *SD* | |
| VR | 73.13 | 77.96 |
| DM | 91.49 | 89.73 |
| QR | 90.73 | 84.76 |
| AR | 98.73 | 97.17 |
| Total Cognitive | 292.79 | 287.20 |
| SJT | 79.04 | 76.08 |

Table 26. Education *T*-Test

| Subtest | *T*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 5.53 | 35,723 | < 0.01 |
| DM | -10.31 | 35,723 | < 0.01 |
| QR | -16.12 | 35,723 | < 0.01 |
| AR | -7.23 | 35,723 | < 0.01 |
| Total Cognitive | -9.22 | 35,723 | < 0.01 |
| SJT | 10.53 | 35,723 | < 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 are resident in the UK, as can be seen in Table 27 below.

Table 27. Candidate Count by Residence

| Country of Permanent Residence | *N* | Percent |
|---|---|---|
| UK | 30,004 | 82% |
| Rest of World | 5,314 | 15% |
| EU | 1,056 | 3% |

In past technical reporting, EU and Rest of World are combined into one category called Non-UK. Since 2017, the proportion of candidates who reside in the UK has slightly increased. However, the proportion has changed between 2021 and 2022, from 83% UK to 82% UK, as shown in Figure 12 below.

Figure 12. Country of Residence 2017–2022

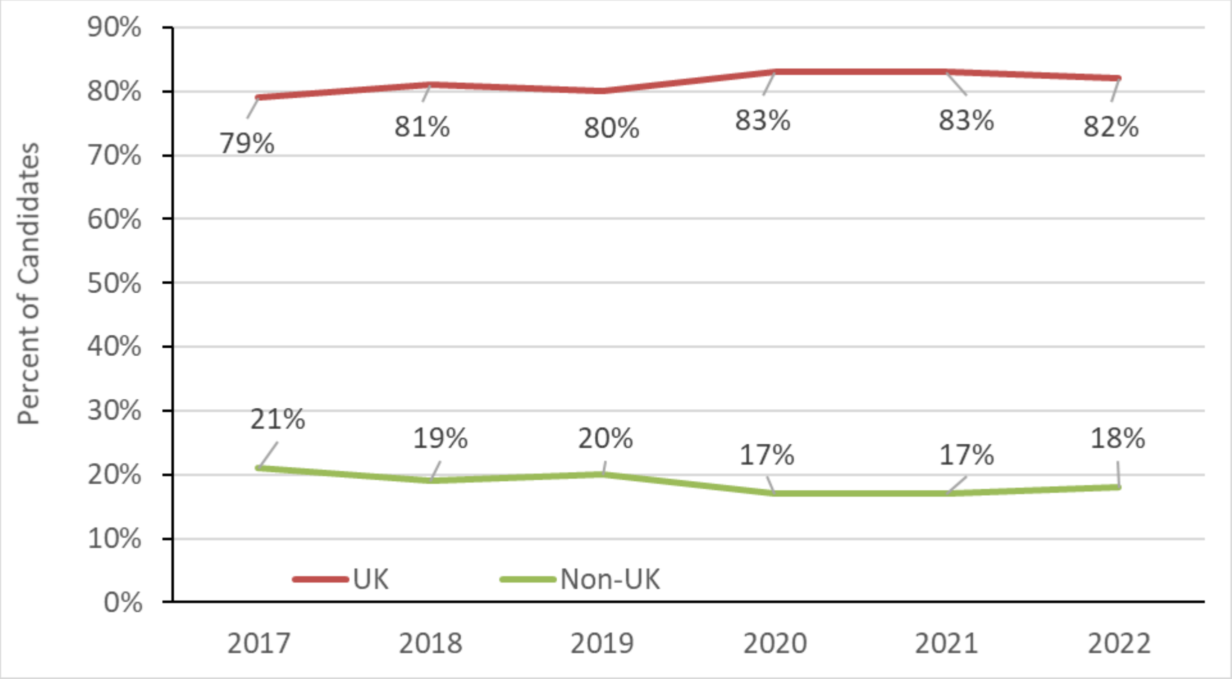


Table 28 shows that UK candidates outperform EU and Rest of World candidates across all subtests.

Table 28. Candidate Scaled Scores by Residence

| Subtest | UK | Rest of World | EU |
|---|---|---|---|
| Mean Scaled Score | | | |
| VR | 570.09 | 551.11 | 557.84 |
| DM | 619.54 | 597.59 | 603.90 |
| QR | 661.49 | 642.85 | 628.85 |
| AR | 664.37 | 635.38 | 639.68 |
| Total Cognitive | 2,515.48 | 2,426.94 | 2,430.27 |
| SJT | 599.59 | 550.71 | 576.26 |
| *SD* | | | |
| VR | 73.00 | 78.72 | 70.81 |

|  |  |  |  |
|---|---|---|---|
| DM | 90.03 | 98.55 | 84.71 |
| QR | 88.40 | 99.61 | 78.86 |
| AR | 96.85 | 105.12 | 98.18 |
| Total Cognitive | 286.30 | 321.29 | 268.37 |
| SJT | 73.65 | 94.65 | 82.75 |

An *F*-test of the differences observed between UK and non-UK candidates is presented in

Table 29 below. It shows that the differences are statistically significant.

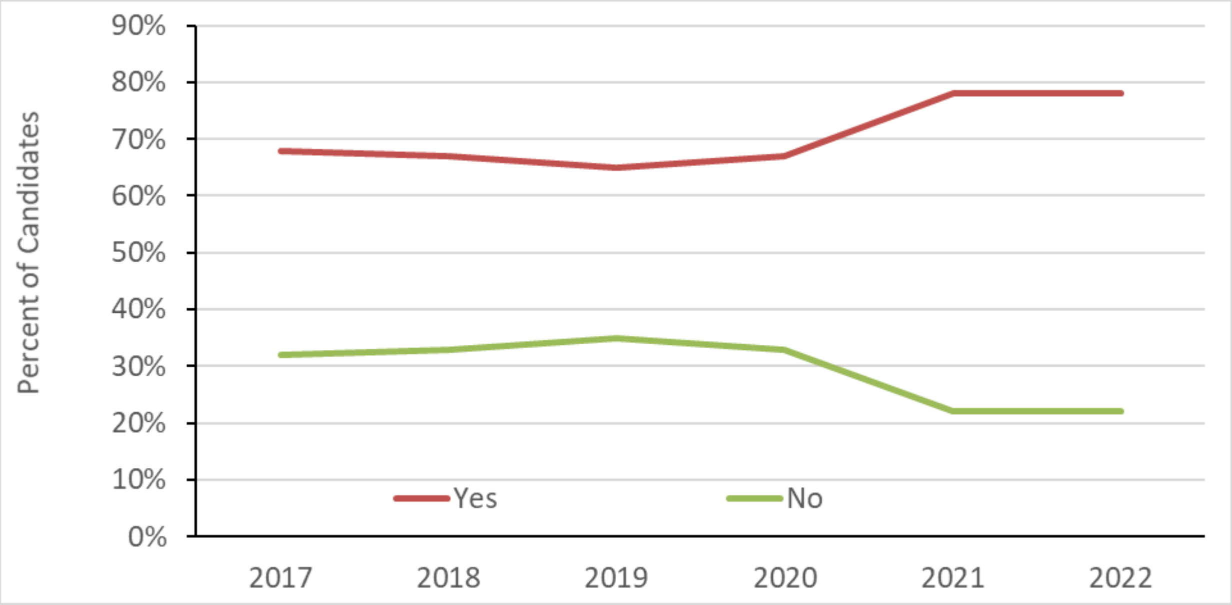
Table 29. Residence *F*-Test

| Subtest | *F*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 157.58 | 2 | < 0.01 |
| DM | 140.19 | 2 | < 0.01 |
| QR | 153.50 | 2 | < 0.01 |
| AR | 218.92 | 2 | < 0.01 |
| Total Cognitive | 239.97 | 2 | < 0.01 |
| SJT | 923.34 | 2 | < 0.01 |

### 4.5.8 First Language

In 2022, most candidates who sat the UCAT stated that English was their first or primary language. Since 2017, the proportion of candidates who state that they speak English as a first or primary language has fluctuated (Figure 13). However, between 2021 and 2022 the proportion of candidates with English as a first language stayed at 78%. The change in 2021 is due to a small change in the wording of this question.

Figure 13. Count of Language 2017–2022



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

Table 30. Scaled Scores by Language and Country of Residence

| Subtest | Country of Residence | First Language | *N* | % of *N* | Mean | *SD* |
|---|---|---|---|---|---|---|
| Verbal Reasoning | UK | English | 25,000 | 69% | 576.88 | 72.24 |
| | | Other | 5,004 | 14% | 536.16 | 67.04 |
| | non-UK | English | 3,404 | 9% | 569.54 | 77.71 |
| | | Other | 2,966 | 8% | 532.35 | 72.33 |
| Quantitative Reasoning | UK | English | 25,000 | 69% | 667.87 | 87.57 |
| | | Other | 5,004 | 14% | 629.58 | 85.6 |
| | non-UK | English | 3,404 | 9% | 650.38 | 94.78 |
| | | Other | 2,966 | 8% | 629.23 | 97.46 |
| Abstract Reasoning | UK | English | 25,000 | 69% | 668.76 | 96.53 |
| | | Other | 5,004 | 14% | 642.42 | 95.43 |
| | non-UK | English | 3,404 | 9% | 640.72 | 99.98 |
| | | Other | 2,966 | 8% | 630.78 | 108.22 |
| Decision Making | UK | English | 25,000 | 69% | 627.94 | 87.9 |
| | | Other | 5,004 | 14% | 577.57 | 88.77 |
| | non-UK | English | 3,404 | 9% | 615.92 | 93.56 |
| | | Other | 2,966 | 8% | 578.81 | 95.84 |
| Total Cognitive Score | UK | English | 25,000 | 69% | 2,541.45 | 280.49 |
| | | Other | 5,004 | 14% | 2,385.72 | 279.66 |
| | non-UK | English | 3,404 | 9% | 2,476.56 | 303.48 |
| | | Other | 2,966 | 8% | 2,371.17 | 314.57 |
| | UK | English | 25,000 | 69% | 604.42 | 69.5 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Situational Judgement |  | Other | 5,004 | 14% | 575.47 | 87.73 |
|  | non-UK | English | 3,404 | 9% | 572.7 | 82.69 |
|  |  | Other | 2,966 | 8% | 534.58 | 100.29 |

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

Table 31. Language *T*-Test

| Subtest | *T*-Statistic | *df* | *p* Value |
|---|---|---|---|
| VR | 45.13 | 36,372 | < 0.01 |
| DM | 42.82 | 36,372 | < 0.01 |
| QR | 32.20 | 36,372 | < 0.01 |
| AR | 21.97 | 36,372 | < 0.01 |
| Total Cognitive | 42.28 | 36,372 | < 0.01 |
| SJT | 41.09 | 36,372 | < 0.01 |

### 4.5.9 Demographic Interactions and SEN

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

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

A shortened version of that analysis was also conducted this year to continue monitoring the differences in the performance between UCAT candidates and UCATSEN candidates, as presented in Table 32. After controlling for the effect of the demographic variables (see the note in Table 32), the difference in exam version still explains a significant amount of variance in the candidates' performance, as candidates who took the UCATSEN performed better than those who took the UCAT. The largest difference was observed in the QR subtest, and the smallest difference was observed in the SJT subtest. The QR and SJT subtests are considered to be the most speeded and least speeded subtests of the exam respectively. This is consistent with our previous hypothesis that the SEN exam advantage is positively associated with the speededness of the exam, and therefore reducing the speededness of the exam should help narrow the performance gap between the candidates who took the accommodated and non-accommodated versions of the exam. The issue of speededness is discussed further in the next section.

Table 32. Subtest Performance Differences: UCAT and UCATSEN

| Subtest | *F* | *p* | $\eta^2$ |
|---|---|---|---|
| VR | 81.91 | <.0001 | .0021 |
| DM | 68.20 | <.0001 | .0018 |
| QR | 120.75 | <.0001 | .0031 |
| AR | 92.30 | <.0001 | .0025 |
| SJT | 9.56 | 0.002 | .0003 |

*Note.* The comparison was only made between UCAT and UCATSEN exam codes, which accounted for 99% of the candidates. The rest of the accommodated exam codes were not included because of the small number of candidates. The demographic variables 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 have not provided 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 are 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

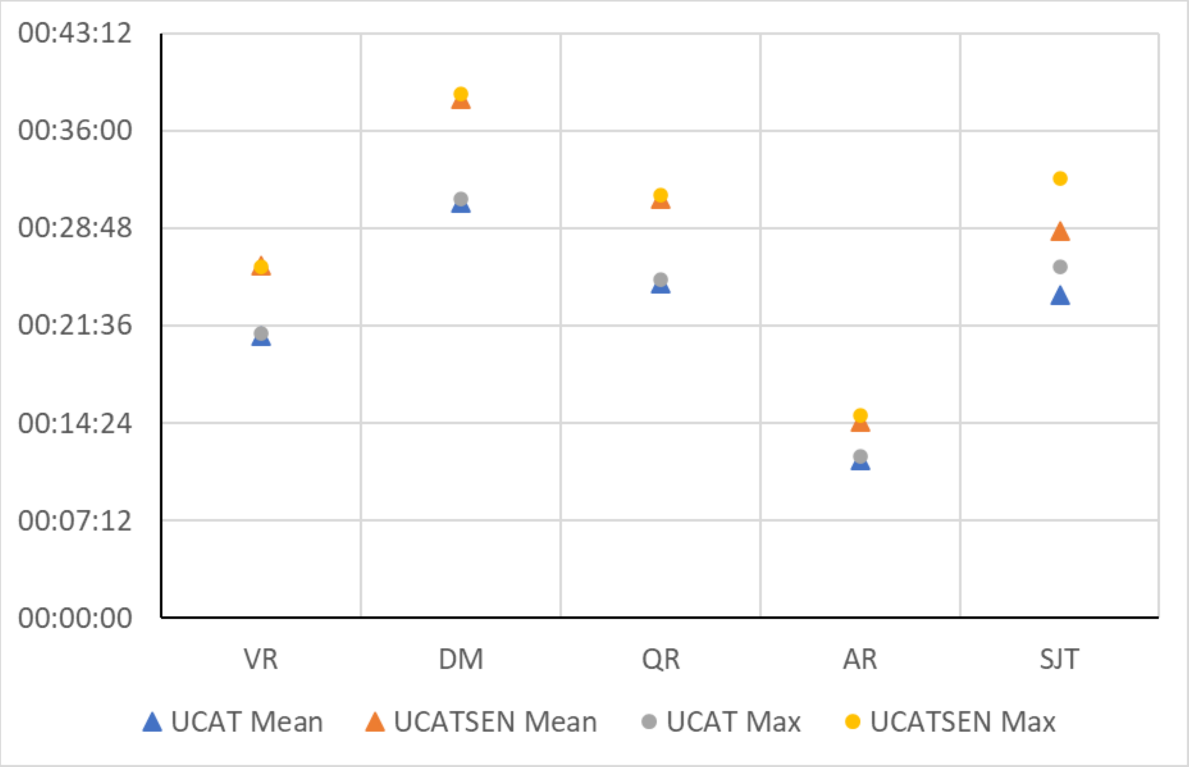
The section time for each candidate is calculated by summing the item and review time for each item and candidate. Table 33 shows the exam timing for each version of the UCAT.

Table 33. Mean Subtest Section Timing: Non-SEN and SEN

| Statistic | Subtest | UCAT (34,181) | UCATSEN (1,605) | UCATSENSA (329) | UCATSEN50 (131) | UCATSA (128) |
|---|---|---|---|---|---|---|
| Mean | VR | 00:20:51 | 00:26:04 | 00:26:06 | 00:31:16 | 00:20:55 |
| | DM | 00:30:43 | 00:38:20 | 00:38:16 | 00:46:02 | 00:30:46 |
| | QR | 00:24:45 | 00:30:57 | 00:30:57 | 00:37:00 | 00:24:52 |
| | AR | 00:11:40 | 00:14:32 | 00:14:32 | 00:17:21 | 00:11:45 |
| | SJT | 00:23:52 | 00:28:39 | 00:28:00 | 00:32:44 | 00:23:37 |
| *SD* | VR | 00:00:30 | 00:00:35 | 00:00:25 | 00:00:43 | 00:00:11 |
| | DM | 00:01:00 | 00:01:38 | 00:01:47 | 00:01:13 | 00:00:32 |
| | QR | 00:01:09 | 00:01:27 | 00:00:56 | 00:01:10 | 00:00:16 |
| | AR | 00:00:47 | 00:01:14 | 00:00:57 | 00:01:21 | 00:00:35 |
| | SJT | 00:03:11 | 00:04:58 | 00:05:17 | 00:06:42 | 00:03:27 |
| Min | VR | 00:01:44 | 00:10:13 | 00:20:27 | 00:25:39 | 00:19:43 |
| | DM | 00:02:17 | 00:08:02 | 00:13:02 | 00:38:26 | 00:27:26 |
| | QR | 00:01:18 | 00:01:20 | 00:21:07 | 00:30:57 | 00:23:02 |
| | AR | 00:01:46 | 00:02:15 | 00:08:25 | 00:08:51 | 00:07:03 |
| | SJT | 00:01:32 | 00:03:19 | 00:12:04 | 00:05:52 | 00:14:06 |
| Max | VR | 00:21:00 | 00:26:16 | 00:26:16 | 00:31:31 | 00:21:01 |
| | DM | 00:31:00 | 00:38:46 | 00:38:46 | 00:46:31 | 00:31:01 |
| | QR | 00:25:00 | 00:31:16 | 00:31:16 | 00:37:31 | 00:25:01 |
| | AR | 00:12:00 | 00:15:01 | 00:15:01 | 00:18:01 | 00:12:01 |
| | SJT | 00:26:00 | 00:32:31 | 00:32:31 | 00:39:01 | 00:26:01 |

There is no agreed definition of speededness, although usually it is assessed by examining how closely the average time candidates spend on a subtest is to the total time allowed, as presented in Table 33. The cognitive subtests on the UCAT version of the exam are quite speeded. The mean time spent completing each subtest is close to the maximum time for each subtest except the SJT, which is considerably less speeded. The SEN versions of the exam are slightly less speeded than the UCAT version. However, the difference between the UCAT version and the UCATSEN version, which is the only SEN version with enough candidates for reliable comparison, is rather small, as shown in Figure 14 below. The difference between the average time and the maximum time allowed is barely observable for VR, DM and QR for both UCAT and UCATSEN. The difference is slightly broader for AR and is quite clear for the SJT.

Figure 14. Mean and Maximum Time for UCAT and UCATSEN



Test timing can be examined in more detail in Table 34. It shows that the most speeded non-SEN subtests are VR and QR, where 87% and 86% of candidates respectively reached all the items and between 6% to 7% of candidates did not reach five or more items. The SJT is the least speeded in all exam versions.

Table 34. Subtest Section Timing: Non-SEN and SEN UCAT Incomplete Tests

| Exam | Subtest | Reached All Items *N* | Reached All Items % | Five or More Items Unreached *N* | Five or More Items Unreached % | Mean Number of Unreached Items for Incomplete Tests Only |
|---|---|---|---|---|---|---|
| UCAT | VR | 29,685 | 87% | 2,308 | 7% | 6.91 (4,496) |
| | DM | 31,615 | 92% | 742 | 2% | 3.61 (2,566) |
| | QR | 29,540 | 86% | 2,136 | 6% | 5.91 (4,641) |
| | AR | 30,757 | 90% | 1,606 | 5% | 6.78 (3,424) |
| | SJT | 33,148 | 97% | 152 | 0% | 3.06 (1,033) |
| UCATSEN | VR | 1,456 | 91% | 67 | 3% | 6.20 (149) |
| | DM | 1,514 | 94% | 23 | 1% | 3.58 (91) |
| | QR | 1,468 | 91% | 63 | 4% | 5.75 (137) |
| | AR | 1,513 | 94% | 42 | 3% | 6.21 (92) |
| | SJT | 1,581 | 100% | 4 | 0% | 3.17 (24) |
| UCATSENSA | VR | 297 | 90% | 12 | 4% | 6.59 (32) |
| | DM | 309 | 94% | 4 | 1% | 3.50 (20) |
| | QR | 304 | 92% | 11 | 3% | 6.48 (25) |

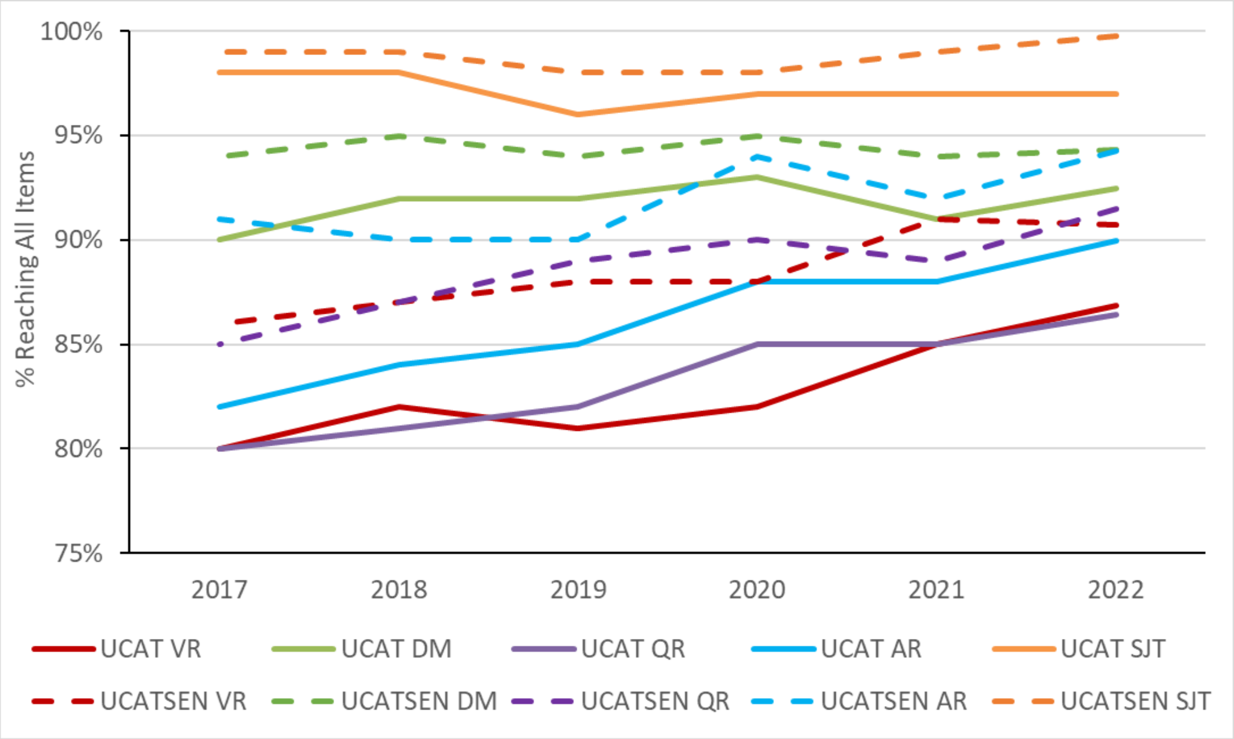
| Exam | Subtest | Reached All Items *N* | Reached All Items % | Five or More Items Unreached *N* | Five or More Items Unreached % | Mean Number of Unreached Items for Incomplete Tests Only |
|---|---|---|---|---|---|---|
| | AR | 312 | 95% | 10 | 3% | 10.53 (17) |
| | SJT | 325 | 99% | 2 | 1% | 13.25 (4) |
| UCATSEN50 | VR | 122 | 93% | 3 | 2% | 5.44 (9) |
| | DM | 128 | 98% | 1 | 1% | 2.33 (3) |
| | QR | 123 | 94% | 2 | 2% | 4.38 (8) |
| | AR | 126 | 96% | 4 | 3% | 7.80 (5) |
| | SJT | 129 | 98% | 0 | 0% | 3.00 (2) |
| UCATSA | VR | 112 | 88% | 6 | 5% | 5.25 (16) |
| | DM | 121 | 95% | 1 | 2% | 2.43 (7) |
| | QR | 115 | 90% | 5 | 4% | 3.69 (13) |
| | AR | 119 | 93% | 4 | 3% | 6.56 (9) |
| | SJT | 126 | 98% | 0 | 0% | 1.00 (2) |

In 2022, a change was made to the timing of the AR and QR subtests with the aim of reducing the speededness of QR. One minute was taken from the AR subtest (with the removal of 5 pretest items) and this was added to the QR subtest (where no additional items were included). The item time has been considered in the form build for QR and AR for a number of years, but this was also extended to VR and DM in 2022. The gaps between the mean time and the total time remained similar to the gaps observed in 2021, with around 10 – 20 seconds difference for all of the cognitive subtests.

The results from 2022 show that the percentage of reached items increased across all four cognitive subtests (and remained at 97% for the SJT). There was a 2% increase in VR and AR and a 1% increase for QR and DM. The increase in reached items for QR is a positive result, and while it is a small increase in the overall number of candidates reaching the end, further analysis showed that there was a decrease in the number of candidates rushing to get to the end.

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

Figure 15. Candidates Reaching All Items 2017–2022



This year an additional approach was adopted to quantify speededness to better monitor the degree of the speeded nature of the test. It was speculated that when candidates perceived they had insufficient time left to complete the test, they would hastily attempt all the remaining items through guessing. Hence, the proportion of the test that candidates were able to attempt before taking a 'guessing' approach could be another way to evaluate the speededness of the test. The guessed attempts were defined as attempts completed in less than a tenth of the average given time per item. For example, VR has 44 items and the time allowed is 21 minutes; on average, each item was allocated 28.63 seconds to complete, and items attempted within 2.86 seconds were defined as guessed responses.

As expected, the proportion of candidates that responded to all of the items dropped after excluding the guessed responses, as shown in Table 35. Consistent with previous analysis, QR was found to be the most speeded subtest, only 23% of candidates responded to all of the items without guessing, and SJT remained the least speeded subtest, up to 74% of candidates responded to all of the items after excluding guessed responses. In both 2021 and 2022, 55% to 65% of candidates responded to all of the items without guessing in the DM and AR sections, only 20% to 30% of candidates do the same for VR and QR, highlighting more substantial speededness issues in VR and QR among the cognitive subtests. Over 90% of candidates were able to respond to more than 85% of the items without guessing in DM, AR and SJT, showing that the subtests provided sufficient time for the majority of candidates to attempt most of the items in these subtests. In contrast, the current test arrangement only ensures that most candidates respond to more than 65% of the items without guessing in VR and QR. When compared to 2021, DM, QR and AR all showed a slight reduction in speededness using this definition, while VR and SJT showed a slight increase in speededness.

Table 35. Proportion of Candidates Who Responded with Non-Guessed Responses

| Percentage of total items | Proportion of candidates who responded without guessing | | | | |
|---|---|---|---|---|---|
| | VR | DM | QR | AR | SJT |
| 2022 | | | | | |
| 100% | 26% | 64% | 23% | 59% | 74% |
| 95% | 43% | 76% | 32% | 77% | 96% |
| 85% | 70% | 92% | 64% | 93% | 99% |
| 75% | 89% | 97% | 83% | 97% | 100% |
| 65% | 96% | 99% | 92% | 99% | 100% |
| 2021 | | | | | |
| 100% | 30% | 59% | 20% | 56% | 77% |
| 95% | 48% | 71% | 29% | 73% | 97% |
| 85% | 74% | 90% | 61% | 91% | 99% |
| 75% | 90% | 97% | 81% | 96% | 100% |
| 65% | 96% | 99% | 90% | 99% | 100% |

*Note*. Only candidates with the “UCAT” exam series code are included in the analysis. Candidates with extra time accommodations were excluded from the analysis. Items that were presented and responded to beyond a tenth of the average given time were considered as non-guessed responses, which included completed items, incomplete items and skipped items.

# 6. Test Form Analysis

The 2022 UCAT consisted of five test forms that were delivered randomly to candidates. Table 36 shows the number of candidates who received each form. Note that SEN candidates were administered Form 1 or Form 2.

Table 36. Candidates by Form

| Form | Candidates |
|---|---|
| Form 1 | 8,060 |
| Form 2 | 7,805 |
| Form 3 | 6,860 |
| Form 4 | 6,739 |
| Form 5 | 6,910 |

Table 37 shows the raw score summary for each subtest on each form. It also includes the reliability statistic, Cronbach's alpha. Alpha is based on the intercorrelations or internal consistency among the items, and it reflects the reproducibility of the test results. High reliability is desirable because it indicates that a test is consistent in measuring the desired construct. AR is consistently the most reliable cognitive subtest, which may be due to the higher number of items. However, all subtests have satisfactorily high reliabilities.

Table 37. Cognitive Raw Score Test Statistics

| Subtest | Form | Mean | *SD* | Min | Max | Alpha | *SEM* |
|---|---|---|---|---|---|---|---|
| VR<br>(40 items) | Form 1 | 21.96 | 5.81 | 3 | 40 | 0.74 | 2.96 |
| | Form 2 | 21.80 | 5.83 | 1 | 40 | 0.75 | 2.92 |
| | Form 3 | 22.05 | 5.78 | 0 | 39 | 0.74 | 2.95 |
| | Form 4 | 21.96 | 5.71 | 2 | 39 | 0.74 | 2.91 |
| | Form 5 | 21.77 | 5.76 | 2 | 39 | 0.74 | 2.94 |
| DM<br>(26 items; 34 score points) | Form 1 | 18.74 | 6.19 | 1 | 34 | 0.80 | 2.77 |
| | Form 2 | 17.20 | 5.81 | 1 | 33 | 0.75 | 2.90 |
| | Form 3 | 17.42 | 5.77 | 2 | 34 | 0.76 | 2.83 |
| | Form 4 | 17.91 | 5.71 | 0 | 33 | 0.77 | 2.74 |
| | Form 5 | 18.84 | 5.82 | 0 | 34 | 0.77 | 2.79 |
| QR<br>(32 items) | Form 1 | 19.54 | 6.29 | 2 | 32 | 0.85 | 2.44 |
| | Form 2 | 19.64 | 6.46 | 1 | 32 | 0.86 | 2.42 |
| | Form 3 | 19.45 | 6.59 | 1 | 32 | 0.86 | 2.47 |
| | Form 4 | 19.21 | 6.05 | 1 | 32 | 0.84 | 2.42 |
| | Form 5 | 19.12 | 6.23 | 2 | 32 | 0.85 | 2.41 |
| AR<br>(50 items) | Form 1 | 33.31 | 8.22 | 5 | 50 | 0.86 | 3.08 |
| | Form 2 | 32.68 | 7.92 | 5 | 50 | 0.85 | 3.07 |
| | Form 3 | 32.30 | 7.97 | 4 | 50 | 0.85 | 3.09 |
| | Form 4 | 32.50 | 8.82 | 5 | 50 | 0.88 | 3.06 |
| | Form 5 | 31.68 | 8.30 | 2 | 50 | 0.86 | 3.11 |

Table 37 also shows *SEM*. This value is the amount of measurement error associated with each subtest and form. *SEM* is calculated using the standard deviation (*SD*) of the raw scores and alpha. Higher reliabilities result in lower *SEM*s.

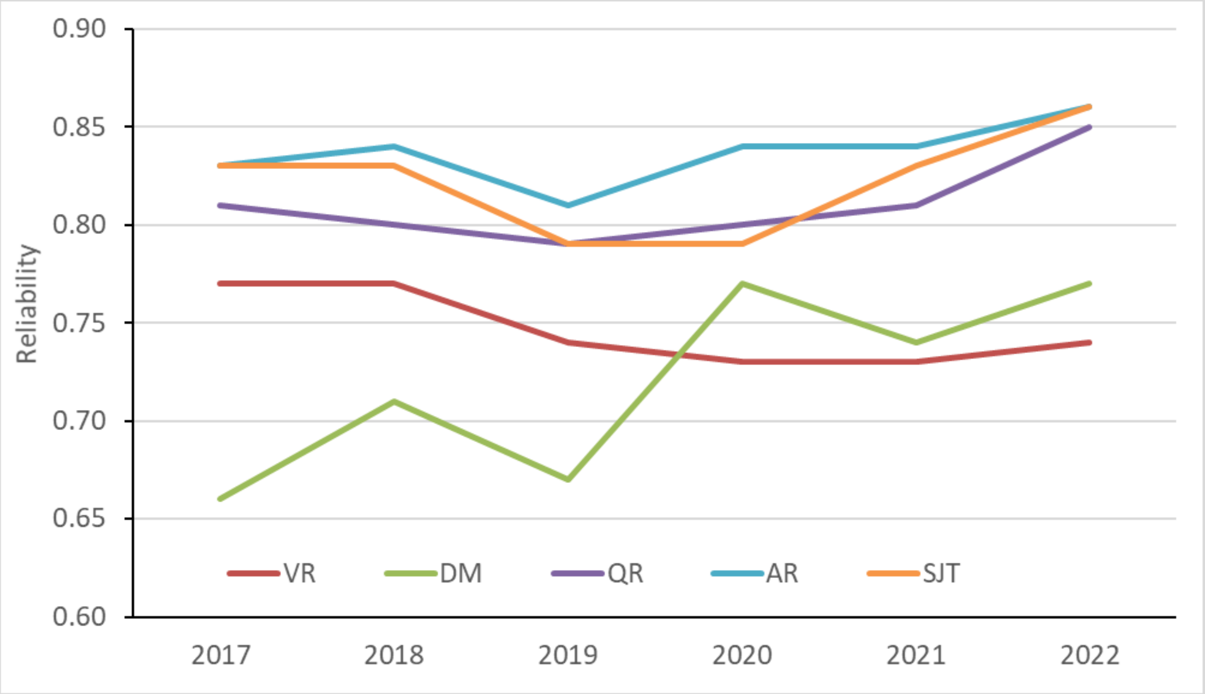
The SJT is analysed in a similar way to the cognitive sections above; however, because the maximum raw score available on the SJT can change year on year, an additional column called mean percent raw score is added (Table 38). Similar to the cognitive results, the reliability is adequately high and the *SEM* adequately low for the SJT.

Table 38. SJT Raw Score Test Statistics (245 score points)

| Form | Mean | *SD* | Min | Max | Mean Percent Raw Score | Alpha | *SEM* |
|---|---|---|---|---|---|---|---|
| Form 1 | 190.61 | 22.46 | 23 | 237 | 77.80% | 0.85 | 8.70 |
| Form 2 | 189.91 | 21.44 | 59 | 237 | 77.53% | 0.84 | 8.58 |
| Form 3 | 189.27 | 24.83 | 39 | 238 | 77.25% | 0.87 | 8.95 |
| Form 4 | 191.95 | 23.72 | 33 | 235 | 78.35% | 0.87 | 8.55 |
| Form 5 | 189.93 | 22.65 | 60 | 239 | 77.52% | 0.85 | 8.77 |

Subtest reliability has been consistent since 2017. Figure 16 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 2022. Interestingly, the reliability of both QR and AR has improved slightly this year. It is possible that this could be attributed to the changes in test timing if it has led to a decrease in guessing towards the end of the test. However, this would need to be monitored over future years to draw any conclusions. There has also been an improvement to the reliability of the SJT, which could be attributed to the introduction of the new ranking item types.

Figure 16. Raw Score Reliability 2017–2021



Raw scores are scaled and reported as scaled scores. The summary statistics for scaled scores on each form are presented below in Table 39. 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 39 also includes the results for the SJT.

Table 39. Cognitive Scaled Score Test Statistics

| Subtest | Form | Mean | *SD* | Min | Max | Reliability | *SEM* |
|---|---|---|---|---|---|---|---|
| VR | Form 1 | 567.33 | 74.51 | 300 | 900 | 0.74 | 37.99 |
| | Form 2 | 566.11 | 75.36 | 300 | 900 | 0.74 | 38.43 |
| | Form 3 | 568.92 | 73.95 | 300 | 890 | 0.73 | 38.43 |
| | Form 4 | 566.51 | 72.90 | 300 | 890 | 0.72 | 38.58 |
| | Form 5 | 565.99 | 73.55 | 300 | 890 | 0.73 | 38.22 |
| DM | Form 1 | 622.06 | 99.04 | 300 | 900 | 0.80 | 44.29 |
| | Form 2 | 608.45 | 86.31 | 300 | 890 | 0.75 | 43.16 |
| | Form 3 | 610.09 | 89.33 | 300 | 900 | 0.77 | 42.84 |
| | Form 4 | 615.28 | 89.22 | 300 | 890 | 0.76 | 43.71 |
| | Form 5 | 623.40 | 91.42 | 300 | 900 | 0.77 | 43.84 |
| QR | Form 1 | 659.59 | 89.80 | 370 | 900 | 0.82 | 38.10 |
| | Form 2 | 661.60 | 92.60 | 310 | 900 | 0.82 | 39.29 |
| | Form 3 | 660.12 | 95.83 | 310 | 900 | 0.83 | 39.51 |
| | Form 4 | 653.72 | 84.11 | 310 | 900 | 0.80 | 37.62 |
| | Form 5 | 653.19 | 87.76 | 370 | 900 | 0.82 | 37.23 |
| AR | Form 1 | 669.09 | 99.70 | 300 | 900 | 0.84 | 39.88 |
| | Form 2 | 659.32 | 94.07 | 300 | 900 | 0.83 | 38.79 |
| | Form 3 | 656.25 | 95.15 | 300 | 900 | 0.84 | 38.06 |
| | Form 4 | 661.51 | 106.57 | 300 | 900 | 0.86 | 39.87 |
| | Form 5 | 649.32 | 97.12 | 300 | 900 | 0.84 | 38.85 |
| Total Cognitive | Form 1 | 2,518.06 | 303.92 | 1,400 | 3,450 | 0.93 | 80.41 |
| | Form 2 | 2,495.49 | 284.53 | 1,280 | 3,470 | 0.92 | 82.95 |
| | Form 3 | 2,495.38 | 293.29 | 1,510 | 3,410 | 0.92 | 82.95 |
| | Form 4 | 2,497.02 | 292.80 | 1,210 | 3,470 | 0.92 | 82.82 |
| | Form 5 | 2,491.90 | 289.00 | 1,460 | 3,480 | 0.92 | 81.74 |
| SJT | Form 1 | 585.73 | 81.34 | 300 | 757 | 0.85 | 31.50 |
| | Form 2 | 591.85 | 77.67 | 300 | 764 | 0.84 | 31.07 |
| | Form 3 | 596.42 | 80.50 | 300 | 757 | 0.87 | 29.02 |
| | Form 4 | 597.62 | 76.78 | 300 | 738 | 0.87 | 27.68 |
| | Form 5 | 588.43 | 79.18 | 300 | 762 | 0.85 | 30.67 |

# 7. Item Analysis

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

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

## 7.1 Cognitive Item Analysis

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

- Point biserial: the degree to which a test item discriminated between strong and weak candidates. For operational items, it must be greater than 0.1 for the item to remain in the bank. For pretest items, it must be greater than 0.05.
- *p* Value: the proportion of candidates who answered the item correctly—the item difficulty. This must be between 0.1 and 0.95 for the item to remain in the bank.
- IRT*b*: the difficulty parameter from the item response theory 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 40 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 40. Cognitive Items Passing the Quality Criteria

| | | VR | | DM | | QR | | AR | |
|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational Scored | Pass | 198 | 99% | 129 | 99% | 160 | 100% | 249 | 100% |
| | Fail | 2 | 1% | 1 | 1% | 0 | 0% | 1 | 0% |
| | *p* < 10 or > 95 | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | pBis <= 0.1 | 2 | 1% | 1 | 1% | 0 | 0% | 1 | 0% |
| | \|b\| >= 3 | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest Unscored | Pass | 251 | 98% | 221 | 96% | 214 | 98% | NA | NA |
| | Fail | 5 | 2% | 9 | 4% | 5 | 2% | NA | NA |
| | *p* < 10 or > 95 | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA |
| | pBis <= 0.05 | 5 | 2% | 9 | 4% | 2 | 1% | NA | NA |
| | \|b\| >= 3 | 0 | 0% | 0 | 0% | 3 | 1% | NA | NA |

Consistent with previous years, only four operational items failed the analysis. Those items did not discriminate highly enough. For the pretest items, few failed in the VR, DM and QR subtests. On VR and DM, the pretest failure was due solely to low item discrimination. For QR, it was due to low discrimination in addition to items being too easy or difficult. There were no pretest items for AR this year. Figure 17 and Figure 18 show that the pretest pass rate has been consistent, with excellent pass rates for VR, DM and QR.

Figure 17. Proportion of Operational Items Failing Analysis 2017–2022

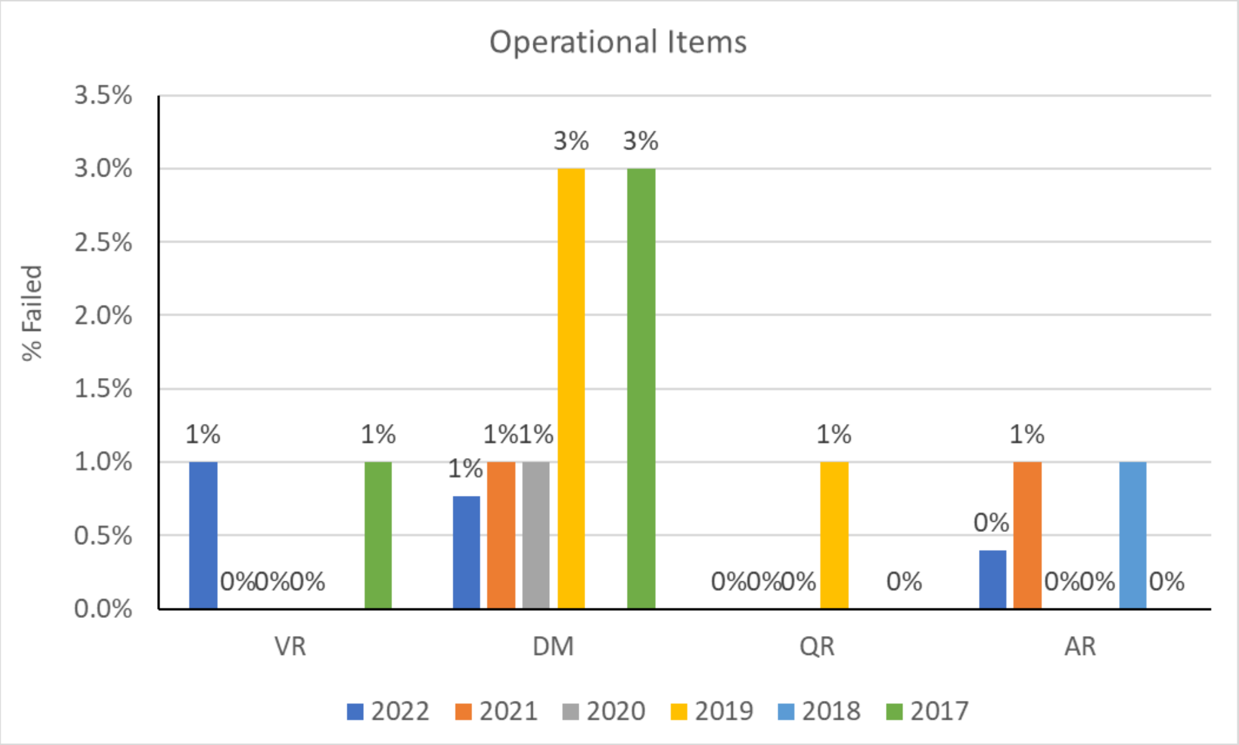


Figure 18. Proportion of Pretest Items Failing Analysis 2017–2022

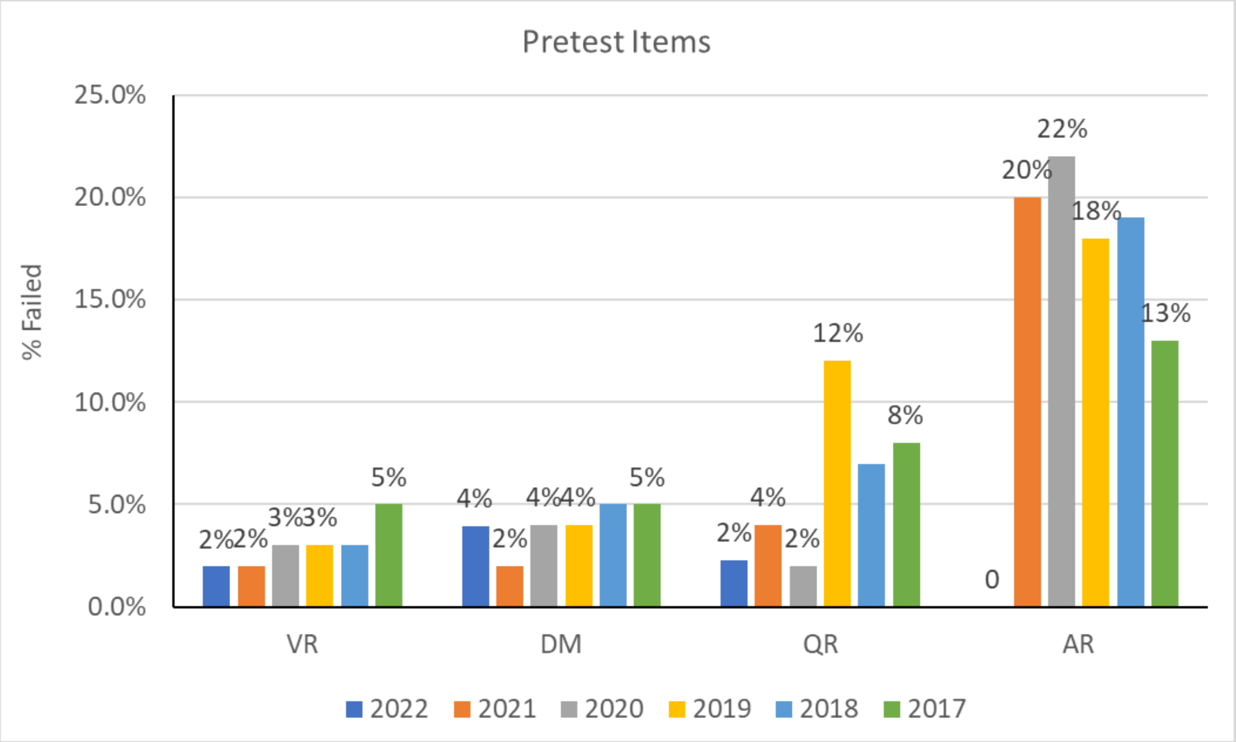


Table 41 shows a summary of the point biserial values. The maximum point biserial is 1, and higher values are better because they indicate that an item can discriminate well between strong and weak candidates. Given that the unscored items have not been tested before, it is expected that those items, on average, will discriminate less well than the scored items, and that is the case across all the cognitive subtests.

Table 41. Discrimination Summary Statistics

| Scored/Unscored | Subtest | *N* Items | Mean pBis | *SD* pBis | Min pBis | Max pBis |
|---|---|---|---|---|---|---|
| Operational (Scored) | VR | 200 | 0.28 | 0.06 | 0.08 | 0.41 |
| | DM | 130 | 0.37 | 0.10 | 0.03 | 0.63 |
| | QR | 160 | 0.40 | 0.07 | 0.13 | 0.57 |
| | AR | 250 | 0.35 | 0.07 | 0.01 | 0.53 |
| Pretest (Unscored) | VR | 256 | 0.24 | 0.08 | -0.08 | 0.42 |
| | DM | 230 | 0.30 | 0.10 | -0.04 | 0.60 |
| | QR | 219 | 0.29 | 0.10 | -0.07 | 0.59 |
| | AR | NA | NA | NA | NA | NA |

Historically, the point biserial values for scored items have been high and stable, whereas the values for unscored items have been lower and less consistent, as illustrated in Figure 19. The operational items appear to have become slightly more discriminating over time for all subtests except VR. This is an indication that the quality of the subtests has improved over time.

Figure 19. Point biserial 2017–2022

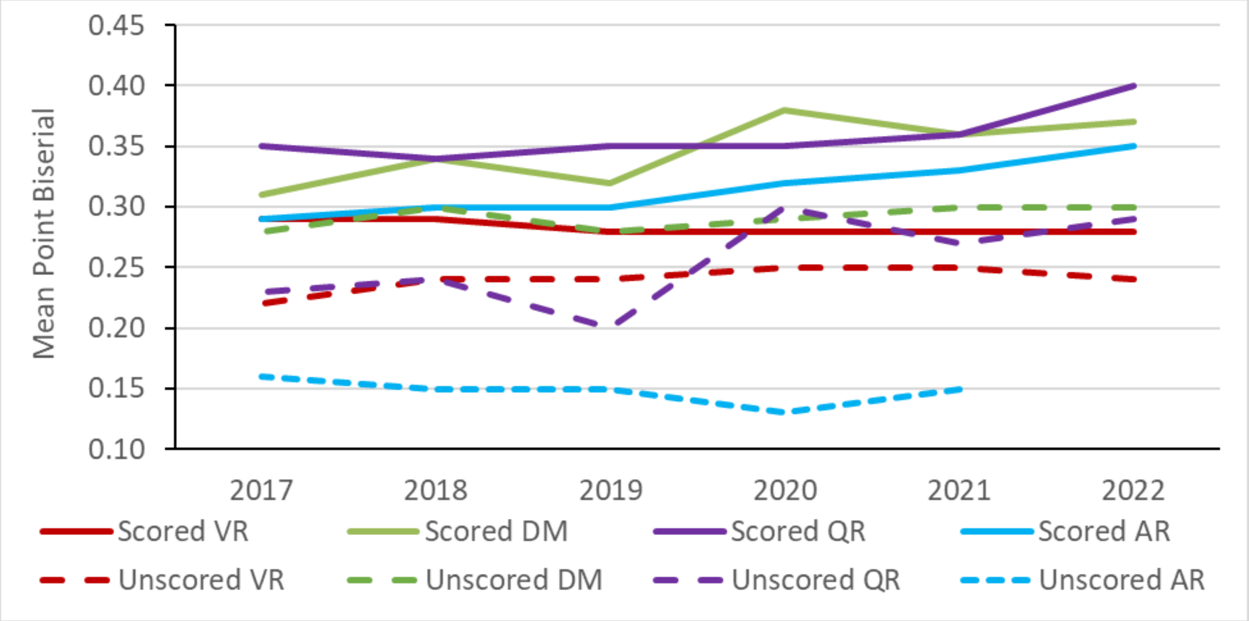


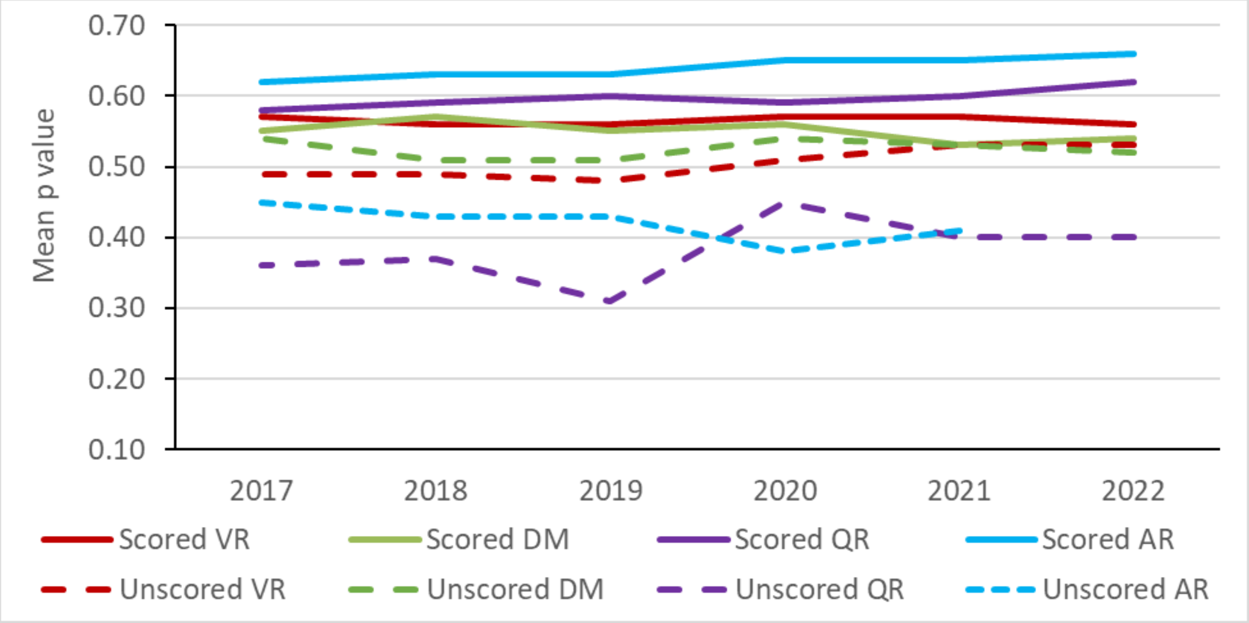
Table 42 shows the summary *of p* values for the cognitive subtests. *p* values reflect the proportion of candidates who answered an item correctly, so higher values indicate easier items, and lower values harder items. Of the operational items, DM items appear to have been the most difficult on average for 2022 candidates and AR items were the easiest on average. The pretest pools appear to have been somewhat more difficult overall than the operational test items for all subtests.

Table 42. *p* Value Summary Statistics

| Scored/Unscored | Subtest | *N* Items | Mean *p* | *SD p* | Min *p* | Max *p* |
|---|---|---|---|---|---|---|
| Operational (Scored) | VR | 200 | 0.56 | 0.13 | 0.12 | 0.91 |
| | DM | 130 | 0.54 | 0.15 | 0.23 | 0.90 |
| | QR | 160 | 0.62 | 0.14 | 0.24 | 0.87 |
| | AR | 250 | 0.66 | 0.14 | 0.24 | 0.91 |
| Pretest (Unscored) | VR | 256 | 0.53 | 0.15 | 0.15 | 0.86 |
| | DM | 230 | 0.52 | 0.19 | 0.12 | 0.94 |
| | QR | 219 | 0.40 | 0.16 | 0.05 | 0.83 |
| | AR | NA | NA | NA | NA | NA |

Since 2017, pretesting has been successful in identifying items that are too difficult and too easy. Figure 20 shows that the items in the pretest pools are usually more difficult than the operational items on average. Note that the subtests are equated year-on-year, meaning changes in difficulty of individual items does not have an impact on the ability required for candidates to achieve a given scaled score.

Figure 20. *p* Value 2017–2022



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

Table 43 shows that the four-option multiple-choice items are better at discriminating between stronger and weaker candidates than the three-option items. The lower point biserials in the pretest pool shows that pretesting is successfully removing items that do not discriminate effectively. The operational items are also rather easier on average than the pretest pool items.

Table 43. 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 | 120 | 0.30 | 0.05 | 0.52 | 0.11 |
| | True/False/Can't Tell | 80 | 0.25 | 0.06 | 0.61 | 0.14 |
| Pretest (Unscored) | Multiple Choice | 136 | 0.26 | 0.07 | 0.51 | 0.14 |
| | True/False/Can't Tell | 120 | 0.21 | 0.09 | 0.56 | 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 44.

Table 44. 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 | 0.47 | 0.07 | 0.52 | 0.14 |
| | Multiple Choice | 90 | 0.33 | 0.09 | 0.55 | 0.16 |
| Pretest (Unscored) | Drag and Drop | 51 | 0.37 | 0.12 | 0.39 | 0.17 |
| | Multiple Choice | 179 | 0.27 | 0.12 | 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 and distinguish slightly better than syllogism items, as presented in Table 45. For the multiple-choice items, the items on statistical reasoning and Venn diagrams are the most discriminating. Interestingly, statistical reasoning was found to be the most difficult item type in DM, while Venn diagrams were found to be the easiest, which shows that discriminating items were included at different levels of difficulty.

Table 45. 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 | 0.45 | 0.06 | 0.48 | 0.12 |
| | | Syllogisms | 20 | 0.49 | 0.08 | 0.55 | 0.16 |
| | Multiple Choice | Logical Puzzles | 20 | 0.3 | 0.08 | 0.48 | 0.13 |
| | | Statistical Reasoning | 20 | 0.38 | 0.06 | 0.47 | 0.14 |
| | | Assumptions Recognition | 20 | 0.25 | 0.08 | 0.58 | 0.12 |
| | | Venn Diagrams | 30 | 0.37 | 0.06 | 0.63 | 0.16 |
| Pretest (Unscored) | Drag and Drop | Information Interpretation | 31 | 0.33 | 0.12 | 0.34 | 0.15 |
| | | Syllogisms | 20 | 0.45 | 0.09 | 0.47 | 0.17 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Multiple Choice | Logical Puzzles | 35 | 0.26 | 0.1 | 0.54 | 0.16 |
| | | Statistical Reasoning | 25 | 0.3 | 0.11 | 0.46 | 0.17 |
| | | Assumptions Recognition | 49 | 0.2 | 0.11 | 0.57 | 0.19 |
| | | Venn Diagrams | 70 | 0.32 | 0.1 | 0.58 | 0.17 |

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 2022 exam, as shown in Table 46.

Table 46. 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 | 0.40 | 0.08 | 0.61 | 0.14 |
| | Standalone | 20 | 0.41 | 0.05 | 0.67 | 0.17 |
| Pretest (Unscored) | Item Set | 198 | 0.29 | 0.10 | 0.39 | 0.15 |
| | Standalone | 21 | 0.33 | 0.12 | 0.50 | 0.19 |

The AR subtest consists of four different types. Table 47 below shows that the discrimination of all four item types is similarly strong across the operational items.

Table 47. 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 | 0.67 | 0.13 |
| | Type 2 | 10 | 0.32 | 0.08 | 0.53 | 0.18 |
| | Type 3 | 15 | 0.28 | 0.09 | 0.62 | 0.22 |
| | Type 4 | 25 | 0.34 | 0.05 | 0.65 | 0.11 |

### 7.1.1 Item Analysis for SEN

An additional analysis was performed this year to examine whether the items perform differently for exams with accommodations. Overall speaking, the item performances did not show substantial differences between the two set of analyses with all of the differences within a third of a standard deviation and most of them within a tenth of a standard deviation, as presented in Table 48. 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 candidate, as reported in previous section. The IRT b value did not show a consistent direction of changes, with some subtests appear to be easier in the SEN exam and some appear more difficult. Most importantly, the items have very close point biserial estimation, indicating that the items were able to retain their ability to discriminate in the SEN examination setting.

Table 48. Item Analysis of UCAT and UCATSEN

| Scored/Unscored | Subtest | Statistics | UCAT | | UCATSEN | |
|---|---|---|---|---|---|---|
| | | | Mean | SD | Mean | SD |
| Operational (Scored) | VR | *p* Value | 0.56 | 0.13 | 0.59 | 0.13 |
| | | Point Biserial | 0.28 | 0.06 | 0.3 | 0.07 |
| | | IRTb | -0.2 | 0.65 | -0.17 | 0.63 |
| | DM | Facility | 0.69 | 0.3 | 0.74 | 0.3 |
| | | Point Biserial | 0.37 | 0.1 | 0.38 | 0.12 |
| | | IRTb | 0.23 | 0.7 | 0.18 | 0.75 |
| | QR | *p* Value | 0.62 | 0.14 | 0.67 | 0.13 |
| | | Point Biserial | 0.4 | 0.07 | 0.4 | 0.08 |
| | | IRTb | -0.27 | 0.72 | -0.25 | 0.72 |
| | AR | *p* Value | 0.66 | 0.14 | 0.7 | 0.13 |
| | | Point Biserial | 0.35 | 0.07 | 0.33 | 0.08 |
| | | IRTb | 0.15 | 0.71 | 0.1 | 0.72 |
| Pretest (Unscored) | VR | *p* Value | 0.53 | 0.15 | 0.56 | 0.19 |
| | | Point Biserial | 0.23 | 0.08 | 0.24 | 0.2 |
| | | IRTb | -0.08 | 0.72 | -0.04 | 1.02 |
| | DM | Facility | 0.6 | 0.24 | 0.63 | 0.29 |
| | | Point Biserial | 0.3 | 0.12 | 0.28 | 0.26 |
| | | IRTb | 0.31 | 0.92 | 0.33 | 1.15 |
| | QR | *p* Value | 0.39 | 0.16 | 0.43 | 0.19 |
| | | Point Biserial | 0.29 | 0.1 | 0.31 | 0.2 |
| | | IRTb | 0.93 | 0.92 | 1.08 | 1.15 |

### 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 2022. This section presents the test items performance across the UK and ANZ population of the 2022 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 49 compares the summary statistics for the operational item analysis for the UCAT 2022 to the UCAT ANZ 2022 values. Across all the subtests, the point biserial summary statistics were similar, with the results from the ANZ population showing slightly higher values, indicating that all operational items discriminated as strongly as expected for the UCAT ANZ population. In terms of the *p* value, which is sample-dependant, the UCAT ANZ population had higher (i.e. easier) average values across subtests. The IRT difficulty, on the other hand, is on a common scale. Table 49 shows that for all subtests, the 2022

UCAT and UCAT ANZ had very similar mean IRT difficulty values, indicating a comparable level of difficulty for both populations.

Table 49. Comparison of Operational Item Statistics: UCAT & UCAT ANZ 2022

| Subtest | Item Statistics | *N* Items | UCAT 2022 | | UCAT ANZ 2022 | |
|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Mean | *SD* |
| VR | *p* Value | 200 | 55.90 | 13.20 | 58.31 | 13.17 |
| | Point biserial | 200 | 0.28 | 0.06 | 0.29 | 0.06 |
| | IRT Difficulty | 200 | -0.20 | 0.65 | -0.20 | 0.64 |
| DM | Facility | 130 | 53.94 | 15.21 | 56.80 | 14.83 |
| | Point biserial | 130 | 0.37 | 0.10 | 0.39 | 0.11 |
| | IRT Difficulty | 130 | 0.23 | 0.70 | 0.25 | 0.67 |
| QR | *p* Value | 160 | 61.96 | 14.46 | 64.50 | 13.26 |
| | Point biserial | 160 | 0.40 | 0.07 | 0.43 | 0.07 |
| | IRT Difficulty | 160 | -0.27 | 0.72 | -0.26 | 0.73 |
| AR | *p* Value | 250 | 65.87 | 13.74 | 65.94 | 12.98 |
| | Point biserial | 250 | 0.35 | 0.07 | 0.38 | 0.08 |
| | IRT Difficulty | 250 | 0.15 | 0.70 | 0.14 | 0.71 |

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 50 summarises the number of items showing drift in the UCAT since 2017 compared to the UCAT ANZ since 2019. Compared to the UCAT 2022, the number of drift items for the UCAT ANZ 2022 is comparable for DM and QR, indicating that the UCAT and UCAT ANZ populations are behaving similarly. For AR, the number of items showing drift in the UCAT ANZ population is lower compared to the UCAT population; for VR, the pattern is reversed, and a higher number of drifted items is observed in the UCAT ANZ population. These items were reviewed by the Content Team and there was no clear explanation for the differences in terms of the cultural sensitivity of the items.

Table 50. Number of Operational Items Showing Drift in UCAT vs UCAT ANZ

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

At present, it is recommended that the degree of drift is monitored in 2023. 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 51. 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 51. Candidate Removal Summary for SJT Item Analysis

| Statistic | Criteria | Number of Candidates Removed |
|---|---|---|
| 1. *Z* score of the scaled score | *Z* score < -4.193 | 0 |
| 2. High number of missing responses | > 1 blank response on operational items | 1,309 |
| 3. Low completion time | Drop in score based on response time | 0 |
| 4. Most/Least items incomplete | Only one response for the most-least items | 381 |

The following item statistics are calculated for the SJT items:

- Item facility: the mean score on the items as a percentage of the maximum score available. It represents the difficulty of the item.
- Item *SD*: the *SD* of the scores on the items. It gives an indication of how well the item is differentiating among candidates.
- Item partial correlation: the correlation of the item score with the total score for the operational items and the scaled score for the pretest items. It compares how individuals perform on a given item with how they perform on the test overall and is a measure of discrimination. Item correlations can be interpreted in the following way:
  - Below 0.13 – poor correlation with the test overall and items within this band are unlikely to be used in an operational test.
  - 0.13 to 0.17 – acceptable correlations. Items within this band will only be included if other items within the scenario have higher item partials.

    - 0.17 to 0.25 – reasonable item performance.
    - Above 0.25 – good item performance.

SJT items should meet the following quality criteria:

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

In 2022, a new ‘dichotomised’ rating item type was introduced within the pretest pool. Candidates are required to decide if a statement relating to a passage is Appropriate/Inappropriate or Important/Not Important, and there are three items associated with each passage. As of January 2023, the scoring method for this new item type is still under review. Under the tentative scoring approach, these items were scored as either 0 or 2 for an incorrect or correct response respectively. The current operational multiple-choice items use a rating system where each of the four options is awarded between 0 and 4. The pretest dichotomous items were a mix of newly written items and items that had been converted from the operational bank. It was expected that converted items would perform better than newly written items as they had performed well prior to conversion.

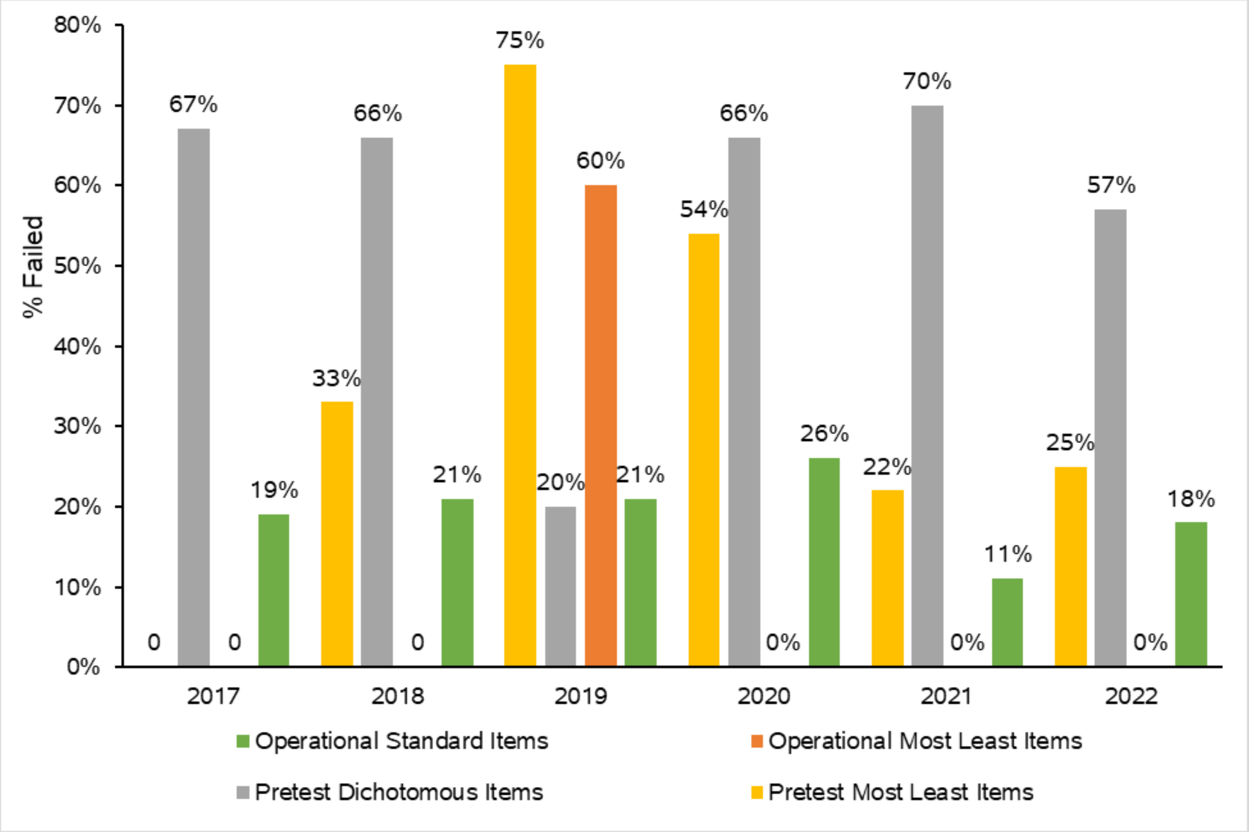
Table 52 shows the number of items that met and did not meet the quality criteria. The most/least item type was more successful than the standard items, with all operational items and 75% of the pretest items meeting the criteria.

Table 52. SJT Items Quality Criteria

|  | Item Type | Statistical Criteria Met/Not Met | All |  | Appropriateness |  | Importance |  | Direct Speech |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | Standard Items | Met | 163 | 82% | 65 | 83% | 75 | 86% | 23 | 70% |
|  |  | Not met | 35 | 18% | 13 | 17% | 12 | 14% | 10 | 30% |
|  | Most/Least Items | Met | 9 | 100% |  |  |  |  |  |  |
|  |  | Not met | 0 | 0% |  |  |  |  |  |  |
| Pretest | Dichotomous Items | Met | 89 | 43% | 30 | 32% | 32 | 58% | 27 | 46% |
|  |  | Not met | 120 | 57% | 65 | 68% | 23 | 42% | 32 | 54% |
|  | Most/Least Items | Met | 21 | 75% |  |  |  |  |  |  |
|  |  | Not met | 7 | 25% |  |  |  |  |  |  |

The proportion of items meeting the quality criteria is fairly consistent with previous years. Figure 21 shows that the proportion of operational standard items not meeting the criteria in 2021 increased from 11% to 18% in 2022. The number of pretest most/least items not meeting the criteria increased slightly from 22% in 2021 to 25% in 2022, however it is important to note that a greater number of most/least items were developed in 2022, from 9 pretest most/least items in 2021 to 29 pretest most/least items 2022, therefore it is positive to see that a high proportion of these items were deemed successful. The new dichotomous items were converted from the previous standard rating items. In 2021, the percentage of standard rating items that met the criteria was 30% and it has increased to 43% of dichotomous items meeting the criteria in 2022, indicating an improvement in pretest item quality. However, it should be noted that a proportion of these items were adapted from already successful items in the item bank and therefore a higher pass rate is expected. Moreover, the item quality statistics are largely dependent on how the items are scored, and since the scoring method of the new dichotomous pretest items is yet to be finalised, the results of these items should be interpreted with caution.

Figure 21. Proportion of SJT Items Failing Analysis 2017–2022



The summary of all operational SJT items is shown below in Table 53.

Table 53. Operational SJT Item Analysis Summary

| Items: 203 | Mean | *SD* | Min | Max |
|---|---|---|---|---|
| Item mean | 3.04 | 1.14 | 0.17 | 7.47 |
| Item *SD* | 1.04 | 0.36 | 0.17 | 2.42 |
| Item partial correlation | 0.28 | 0.11 | -0.02 | 0.53 |
| Item total facility | 75% | 19% | 6% | 100% |

Since 2017, the item mean score and facility has tended to increase, as illustrated in Figure 22, indicating that items are becoming somewhat easier. The total number of items has also progressively increased since 2017. The increase in item partial correlation indicates that despite the test being relatively easy, it has progressive improvement in consistently measuring the same ability and the items are getting better overall at discriminating among strong and weak candidates. A better discrimination between candidates implies that the test results could be considered as being more reliable in distinguishing stronger and weaker candidates. In other words, improvement is seen in the item quality. However, a relatively high facility could imply that the test might be too easy to distinguish between strong and very strong candidates. In future test development, harder items should be developed to minimise the upward trend of item facility.

Figure 22. SJT Operational Summary 2017–2022

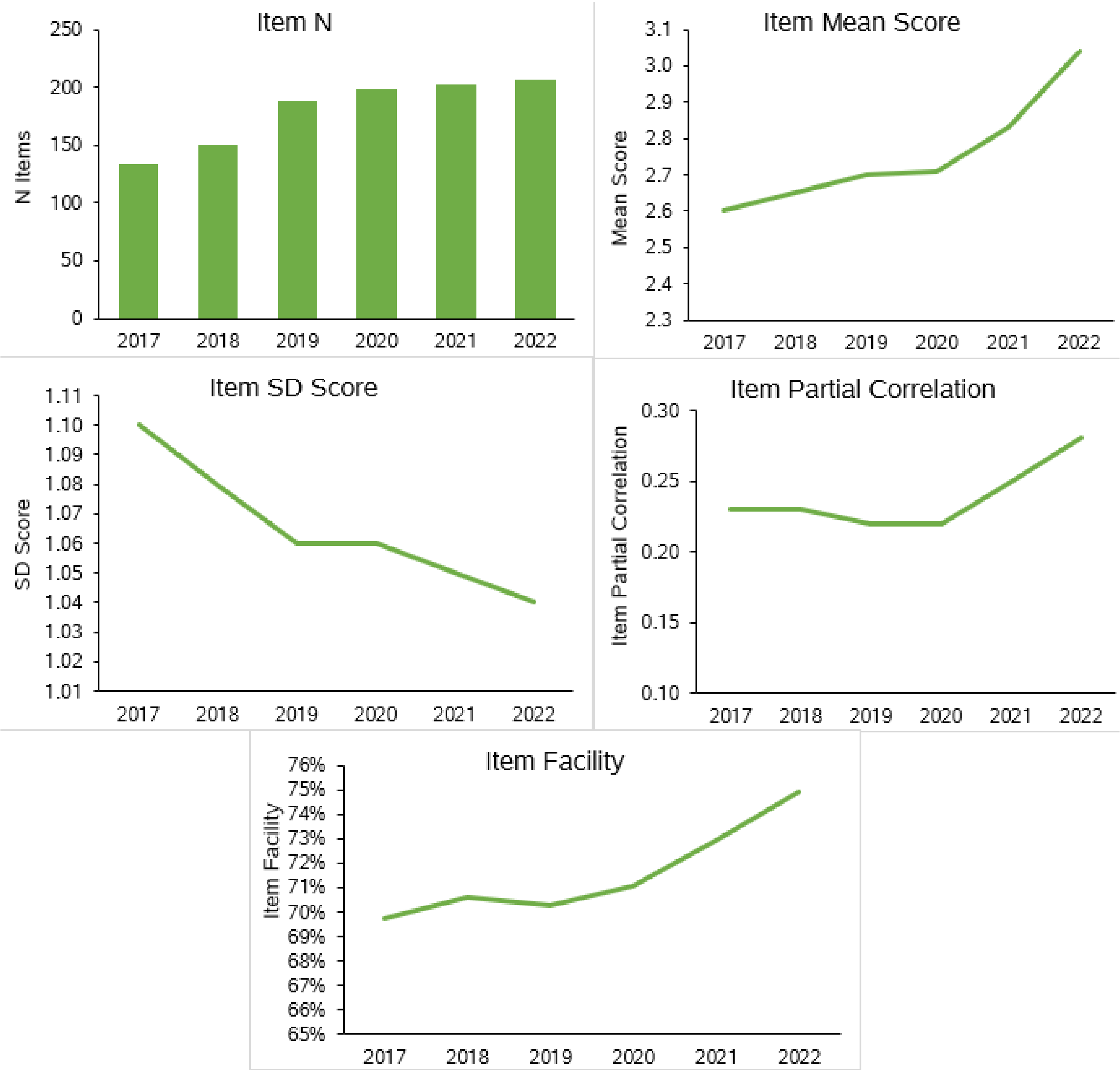


Table 54 shows the summary statistics for the SJT pretest items. In 2022, a new dichotomous item type was pretested in the SJT. A proportion of these items were newly written for the pretest pool, and a proportion were adapted from existing successful scenarios in the item bank. These items performed comparably to the standard items.

Table 54. SJT Pretest Item Summary Statistics

| | Statistic | Item Mean | Item *SD* | Item Partial | Item Total Facility |
|---|---|---|---|---|---|
| Dichotomous Items (209 items) | Mean | 1.50 | 0.65 | 0.15 | 75% |
| | *SD* | 0.50 | 0.28 | 0.10 | 25% |
| | Min | 0.13 | 0.00 | -0.11 | 6% |
| | Max | 2.00 | 1.00 | 0.42 | 100% |
| Most/Least (28 items) | Mean | 6.79 | 1.70 | 0.19 | 85% |
| | *SD* | 0.72 | 0.50 | 0.08 | 9% |
| | Min | 5.20 | 0.86 | 0.01 | 65% |
| | Max | 7.73 | 2.74 | 0.35 | 97% |

## 7.3 Differential Item Functioning

### 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 2022 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 now reported for each demographic group. Table 55 below shows DIF in relation to gender. No items were found to exhibit Category C DIF in any of the subtests.

Table 55. Gender DIF

| Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 200 | 100% | 126 | 97% | 160 | 100% | 250 | 100% | 202 | 98% |
| | B | 0 | 0% | 4 | 3% | 0 | 0% | 0 | 0% | 5 | 2% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | A | 255 | 100% | 230 | 100% | 219 | 100% | NA | NA | 226 | 95% |
| | B | 1 | 0% | 0 | 0% | 0 | 0% | NA | NA | 11 | 5% |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |

In contrast to previous years, the age comparison has been changed to increase the number of items where a comparison can be made. In 2022, the comparison is between less than 20 and greater than 25 in contrast to less than 20 and greater than 35 in previous years (Table 56). Five Category C items were identified in DM, with one item favouring older candidates and four favouring younger candidates. One VR item showed Category C DIF, and this item favoured younger candidates over older candidates. These items will be reviewed by the content team and removed from the bank.

Table 56. Age DIF

| Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 195 | 98% | 112 | 86% | 159 | 99% | 247 | 99% | 204 | 99% |
| | B | 4 | 2% | 13 | 10% | 1 | 1% | 3 | 1% | 3 | 1% |
| | C | 1 | 1% | 5 | 4% | 0 | 0% | 0 | 0% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | A | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 227 | 96% |
| | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 10 | 4% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | NA | 256 | 100% | 230 | 100% | 219 | 100% | NA | NA | 0 | 0% |

For ethnicity, there were usually enough items to reliably categorise DIF for operational items. However, because there were more pretest items, many of the pretest comparisons are not possible due to low candidate numbers. Note that the options on the ethnicity question have changed in 2022 and the “UK-Chinese” category is no longer separate. In addition, a comparison between White and Non-White was included.

Table 57 shows there was one instance of Category C DIF identified in the ethnicity comparisons. The item appeared in the SJT and favoured White candidates over Mixed candidates.

Table 57. Ethnicity DIF

| Type | Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | White/ Black | A | 199 | 99% | 126 | 97% | 157 | 98% | 247 | 99% | 194 | 94% |
| | | B | 1 | 1% | 4 | 3% | 3 | 2% | 3 | 1% | 13 | 6% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | White/ Asian | A | 198 | 98% | 124 | 95% | 156 | 98% | 250 | 100% | 196 | 95% |
| | | B | 2 | 2% | 6 | 5% | 4 | 2% | 0 | 0% | 11 | 5% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | White/ Mixed | A | 200 | 100% | 129 | 99% | 160 | 100% | 249 | 100% | 207 | 100% |
| | | B | 0 | 0% | 1 | 1% | 0 | 0% | 1 | 0% | 0 | 0% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | White/ Non-White | A | 199 | 100% | 127 | 98% | 159 | 99% | 250 | 100% | 193 | 93% |
| | | B | 1 | 0% | 3 | 2% | 1 | 1% | 0 | 0% | 14 | 7% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | White/ Black | A | 103 | 40% | 40 | 17% | 192 | 88% | NA | NA | 218 | 92% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 14 | 6% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | | NA | 153 | 60% | 190 | 83% | 27 | 12% | NA | NA | 5 | 2% |
| | White/ Asian | A | 256 | 100% | 225 | 98% | 219 | 100% | NA | NA | 220 | 93% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 17 | 7% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | | NA | 0 | 0% | 5 | 2% | 0 | 0% | NA | NA | 0 | 0% |
| | White/ Mixed | A | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 218 | 92% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 7 | 3% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 1 | 0% |
| | | NA | 256 | 100% | 230 | 100% | 219 | 100% | NA | NA | 11 | 5% |
| | White/ Non-White | A | 256 | 100% | 229 | 100% | 219 | 100% | NA | NA | 225 | 95% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 12 | 5% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | | NA | 0 | 0% | 1 | 0% | 0 | 0% | NA | NA | 0 | 0% |

In addition to the comparisons produced in previous years, a comparison between SEC1 and non-SEC-1 was also included to allow more comparisons to be made. One Category C DIF item was identified in the SEC comparisons for the operational items, and this item was in the QR subtest and favoured SEC1 over SEC2. As Table 58 demonstrates, very few comparisons were possible for the pretest items. In the SJT, one item favoured SEC 1 over SEC 4.

Table 58. SEC DIF

| Type | Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | SEC 1/2 | A | 200 | 100% | 130 | 100% | 160 | 100% | 250 | 100% | 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% |
| | 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 | 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% |
| | SEC 1/5 | 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/(2-5) | A | 200 | 100% | 130 | 100% | 159 | 99% | 250 | 100% | 207 | 100% |
| | | B | 0 | 0% | 0 | 0% | 1 | 1% | 0 | 0% | 0 | 0% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | SEC 1/2 | A | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 230 | 97% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 4 | 2% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | | NA | 256 | 100% | 230 | 100% | 219 | 100% | NA | NA | 3 | 1% |
| | SEC 1/3 | A | 0 | 0% | 11 | 5% | 2 | 1% | NA | NA | 235 | 99% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 2 | 1% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | | NA | 256 | 100% | 219 | 95% | 217 | 99% | NA | NA | 0 | 0% |
| | SEC 1/4 | A | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 232 | 98% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 1 | 0% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 1 | 0% |
| | | NA | 256 | 100% | 230 | 100% | 219 | 100% | NA | NA | 3 | 1% |
| | SEC 1/5 | A | 1 | 0% | 14 | 6% | 11 | 5% | NA | NA | 232 | 98% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 5 | 2% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | | NA | 255 | 99% | 216 | 94% | 208 | 95% | NA | NA | 0 | 0% |
| | SEC 1/(2-5) | A | 256 | 100% | 190 | 83% | 219 | 100% | NA | NA | 235 | 99% |
| | | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 2 | 1% |
| | | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | | NA | 0 | 0% | 40 | 17% | 0 | 0% | NA | NA | 0 | 0% |

As Table 59 illustrates, there was one Category C DIF item detected in the comparison between candidates who had an honours degree or above and those who did not. This item was a QR pretest item and favoured candidates with a degree education over those without. There were high candidate volumes across the board, meaning comparisons could be made for all subtests.

Table 59. Honours Degree DIF

| Type | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 200 | 100% | 127 | 98% | 159 | 99% | 250 | 100% | 206 | 100% |
| | B | 0 | 0% | 3 | 2% | 1 | 1% | 0 | 0% | 1 | 0% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | A | 256 | 100% | 213 | 93% | 218 | 100% | NA | NA | 226 | 95% |
| | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 11 | 5% |
| | C | 0 | 0% | 0 | 0% | 1 | 1% | NA | NA | 0 | 0% |
| | NA | 0 | 0% | 17 | 7% | 0 | 0% | NA | NA | 0 | 0% |

Comparison was also possible for the most part across all subtests for candidates who reported English as being their first or primary language and those who reported that it was not. As Table 60 shows, no item was flagged as having Category C DIF.

Table 60. English as First Language DIF

| Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 198 | 99% | 130 | 100% | 160 | 100% | 250 | 100% | 206 | 100% |
| | B | 2 | 1% | 0 | 0% | 0 | 0% | 0 | 0% | 1 | 0% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | A | 256 | 100% | 230 | 100% | 219 | 100% | NA | NA | 226 | 95% |
| | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 11 | 5% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |

No Category C DIF items were identified in the comparison of candidates who reported the UK as their residence with those who reported the UK as not being their residence (as presented in Table 61).

Table 61. Residency DIF

| Group | Code | VR | | DM | | QR | | AR | | SJT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % | *N* | % | *N* | % |
| Operational | A | 197 | 98% | 129 | 99% | 153 | 96% | 249 | 100% | 203 | 98% |
| | B | 3 | 2% | 1 | 1% | 7 | 4% | 1 | 0% | 4 | 2% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | NA | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Pretest | A | 256 | 100% | 228 | 99% | 219 | 100% | NA | NA | 231 | 97% |
| | B | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 6 | 3% |
| | C | 0 | 0% | 0 | 0% | 0 | 0% | NA | NA | 0 | 0% |
| | NA | 0 | 0% | 2 | 1% | 0 | 0% | NA | NA | 0 | 0% |

Very few candidates took the online version of the UCAT (231 candidates, see Section 3.4), so comparison was not possible.

In summary, 10 items were found to exhibit Category C DIF. These items were removed from the bank so they will not be selected for future forms.

# 8. Summary

The scores in the 2022 administration of the UCAT were broadly in line with scores in previous years. The proportion of candidates taking a SEN version of the exam did not change, nor did the demographic makeup of test-takers.

Candidates taking a SEN version of the exam continue to score better than candidates taking the non-SEN version, and demographic trends in scores and candidate volumes were consistent with previous years' administrations of the exam. Higher scores continue to be associated with candidates who are resident in the UK, White ethnicity, in SEC 1, and speak English as a first language. Certain scoring patterns by demographic also persist in the 2022 version of the exam. Male candidates outperformed female candidates on the cognitive sections and *vice versa* on the SJT, and candidates with a degree or who were aged 20–24 performed better on the VR and SJT subtests, whereas candidates without a degree or who were aged 16–19 performed better on the other subtests.

The 2022 exam does appear to be slightly less speeded than previous versions of the exam following the changes to the timings of the AR and QR subtests. Several subtests appear to have become somewhat less speeded over the last five years.

In terms of test quality, the test forms were reliable, with appropriately low measurement error, and individual items performed well, with very few operational cognitive items needing to be retired. More SJT items needed to be removed, but that is consistent with performance in previous years.

## 8.1 Recommendations

The outcome of the UCAT 2022 analysis identifies certain small operational changes that have improved the ongoing performance of the test, as well as several areas that might provide fertile ground for further research.

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

The speededness of the cognitive subtests has fallen slightly on some subtests given the changes made in 2022, but the degree of speededness is still potentially a concern. Currently Pearson VUE constructs test forms with a constraint on the historic average response time for all of the subtests to minimise this. This constraint means items that take a long time to answer are not included in the forms. As a result, the average time to answer all items on the subtest is kept to a reasonable level, and this will continue to be monitored. No changes to the test timings are recommended for 2023 at present.

Following the continued development of dichotomous items in the SJT, it is recommended that this item type continues to be developed and used within future pretests. This will allow for further psychometric testing in addition to building a bank of dichotomous items for future use.