University Admissions Tests UK

**UAT-UK Ltd**

# Annual Reports 2024-2025

## TMUA Technical Report

Published September 2025

## Contents

## 1. Executive Summary

This report is based on analysis conducted by UAT-UK's delivery partner, Pearson VUE, as part of their annual process of monitoring and evaluation.

The Test of Mathematics for University Admission (TMUA) was administered in two windows: 16th and 17th October 2025, and 8th and 9th January 2025, for candidates applying to start university in 2025. This report covers the 13,855 candidates (8,987 in October 2024 and 4,868 in January 2025) who tested during the two events.

When candidates register for the test, they complete a questionnaire about their demographic characteristics such as gender, ethnicity and school type. The scaled score patterns split by these different demographic variables generally follow national trends for mathematics-based exams. On average, male candidates tend to outperform female candidates, and candidates with higher socio-economic status generally perform better than those with widening participation flags. Candidates who stated that their first language was not English outperformed native English-language speakers due to a strong performance from candidates outside the UK.

Multiple forms (versions) of the TMUA were used, and these were administered at different times to different regions. The reliabilities for the forms were good, and the corresponding standard errors of measurement (*SEM*s) were low. The forms were very well balanced in terms of difficulty.

The candidates taking TMUA include very able students, so the test is designed to challenge even the best candidates and includes difficult questions. Given this, it is expected that the test will be slightly speeded for the majority of candidates. Most candidates used the full time available for the test, as the mean test time was close to the full time and a small proportion of candidates did not reach the last item in each timed section.

The items used in the TMUA were all new and had not been previously pretested. Despite this, the items performed very well and showed a good range of item difficulties, as is desirable for an admissions test. In addition, the items had a high mean point biserial, indicating that they are generally discriminating well. Relatively few items were flagged as showing DIF, which is an indicator of possible bias.

In conclusion, the results of the 2024/25 TMUA administration were excellent. The test demonstrated good reliability, low measurement error, and balanced difficulties across the forms.

## 2. Introduction

The TMUA is designed to support universities in identifying strong applicants for mathematics-related degree courses. It is used by a number of universities to differentiate among a large number of strong candidates with similar academic profiles.

The TMUA is available in two sittings per admissions cycle. This report covers those candidates who took the test in October 2024 (16th and 17th) and January 2025 (8th and 9th). The test consists of two separately timed sections, and candidates receive a scaled score from 1.0 to 9.0 after the results are processed.

Section 3 of this report outlines the structure of the test and the measurement approach taken for it; Section 4 describes the test results including the overall scaled score results, proportion of candidates requiring accommodations and candidate demographic characteristics.

Following the analysis of results by demographic, the test level performance is summarised in Section 5. This includes the reliability and standard error of measurement (*SEM*), and an analysis of test timing, speededness and unreached items.

The final analysis section, Section 6, summarises item performance across the test as well as a differential item functioning (DIF) analysis by the demographic variables where there were sufficient candidates.

## 3. Test Design and Measurement Approach

### 3.1 Test Design

The 2024/2025 TMUA contains two timed sections (Table 1): Paper 1 and Paper 2 (this is in reference to the previous paper-based version of the test). A number of forms, or versions, of the test are available during the window to allow candidates to test over multiple days. Every effort is made to ensure that these forms are as comparable as possible in terms of content and difficulty to make sure the test is as fair as possible.

Table 1 TMUA Test Design

| Test | Section | Content | Questions | Duration |
|---|---|---|---|---|
| TMUA | Paper 1: Applications of Mathematical Knowledge | Assesses candidates' ability to apply knowledge of mathematics in new situations. | 20 multiple-choice questions | 75 minutes |
| | Paper 2: Mathematical Reasoning | Assesses candidates' ability to deal with mathematical reasoning and simple ideas from elementary logic. | 20 multiple-choice questions | 75 minutes |

Candidates are given 150 minutes to answer a total of 40 items. Candidates are also able to apply for extra time accommodations if required.

Candidates are awarded a scaled score from 1.0 to 9.0 which is reported to one decimal place. Unlike a raw score, which is a function of the candidate ability and test form difficulty, the scaled score is on a single scale and is comparable within this admissions cycle, regardless of the form that was taken. Therefore, a candidate who scored 6.5, for example, in either January or October has a higher ability than a candidate who scored 4.2 in either event. Further details on the scoring process are provided in the subsequent sections.

### 3.2 Measurement Model

#### 3.2.1 Item Analysis

For the 2024/2025 cycle, none of the items in the TMUA had been pretested and therefore they did not have statistics to help guide the selection process. Items were selected for the forms by the Chair based on their expert judgement.

Items are calibrated using an item response theory (IRT) model at the end of each event window. IRT is a theoretical framework that models test responses resulting from an interaction between candidates and test items. The advantage of using IRT models in scaling is that all items measuring performance in one latent trait can be placed on the same scale of difficulty, set using the initial

item analysis. Placing items on the same scale across years facilitates the creation of equivalent forms each year.

For TMUA, the Rasch IRT model was used for item calibration using Winsteps software (Linacre, 2014). Under the Rasch IRT model, the probability of a candidate answering an item correctly is a function of the item difficulty and the candidate's ability. As a candidate's ability increases, his or her chance of correctly answering the item also increases. Mathematically, the probability of candidate $j$ answering item $i$ correctly is defined as:

$$P_{ij} = \frac{\exp(\theta_j - b_i)}{1 - \exp(\theta_j - b_i)}$$

Candidate ability is represented by the variable $\theta$ (theta) and item difficulty (also called the *b* value) by the model parameter $b$. Both $\theta$ and $b$ are expressed on the same metric, with greater values representing either greater ability or greater item difficulty, respectively.

During the item calibration process, the parameters (that is, item difficulty and candidate ability) are not fixed. This is known as scale indeterminacy. However, items can be anchored at their known difficulty values, allowing new item difficulty values to be estimated relative to these fixed values on a common scale.

### 3.2.2 Equating and Scaling

The raw score a candidate achieves on a test is a function of both candidate ability and the item difficulties on the form. If there are multiple forms of a test, this can lead to small differences in difficulty across the forms, despite the best attempts of the Chairs to make the forms comparable when they are put together. In order to treat candidates fairly, these difficulty differences are removed through equating, which places all candidates onto a single ability (or theta) scale, regardless of the form they took. The theta estimate for each candidate is then scaled to generate an easily interpretable score for the candidate. The scaled scores issued to candidates are therefore on a single scale within each admissions cycle and can be used to compare candidates, which is the prime objective of an admissions test. For TMUA, the candidates are issued a scaled score ranging from 1.0 to 9.0 reported to one decimal place. This is consistent with the previous paper version of the test.

TMUA is post-equated, which means that the equating is conducted at the end of the testing window. Therefore, candidates do not receive an immediate score. This has many advantages, including allowing the use of un-pretested operational items in the test and being able to generate a scaled score based on the observed candidate population as opposed to a benchmark population.

Following item analysis, the item difficulties are used to generate a raw score to theta (or ability) table. The theta value is then scaled to generate the scaled score from 1.0 to 9.0. University Admissions Tests UK (UAT-UK) requested that the scaling approach be fixed to the candidate ability

distribution. After the initial analysis of the October 2024 data, it was determined that the median candidate theta should be fixed to a scaled score of 4.5 and the candidate ability corresponding to the 90th percentile should be fixed to a scaled score of 7.0 (Table 2). A regression line was then plotted between these two points to determine the scaling constants (Table 3) used to transform the theta values to scaled scores, which were capped at 1.0 and 9.0. The same scaling constants were used for both the October 2024 and January 2025 events to ensure the scaling was consistent and scaled scores were comparable across events.

Table 2 Ability Estimates Used to Scale the TMUA

| Test | Percentile | Ability | Scaled Score |
|---|---|---|---|
| TMUA (Oct 2024) | 50 | 0.0057 | 4.5 |
| | 90 | 1.3947 | 7.0 |

Table 3 Scaling Constants

| Test | Constant | Multiplier |
|---|---|---|
| TMUA (Oct 2024) | 4.4897 | 1.7998 |

# 4. Test Results

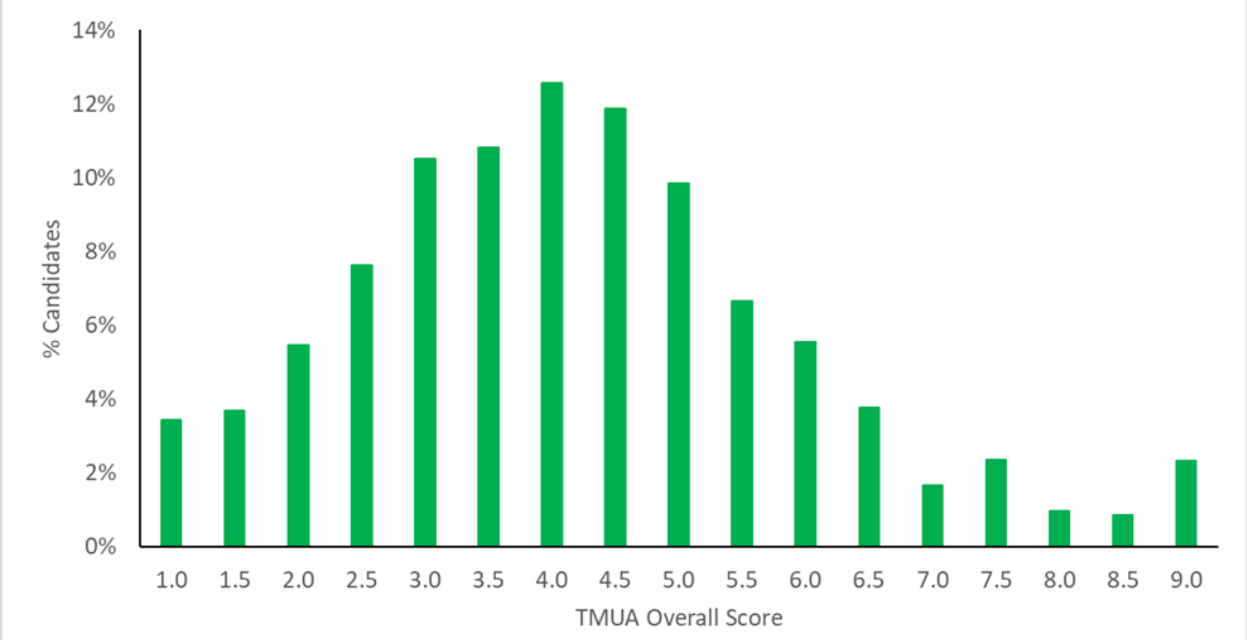
## 4.1 Candidate Performance

This report covers test results for the 2025 admissions cycle, which includes the October 2024 event and the January 2025 event. There were 13,855 candidates in total, with 8,987 (65%) sitting in October 2024 and 4,868 (35%) in the January 2025 event (Table 4). Candidates are only allowed to sit the test once within each admissions cycle and those applying to the University of Cambridge must sit in the October event, which accounts for the higher volume in this session. The scaled score statistics for the complete cohort and each event are summarised in Table 4. The scaled score distribution is illustrated in Figure 1. Candidate scaled scores are well distributed across the scaled score range, enabling universities to effectively differentiate between candidates.

Table 4 Scaled Score Summary Statistics

| Event | *N* | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| All | 13,855 | 4.20 | 1.77 | 1.0 | 9.0 | 2.9 | 4.0 | 5.2 | 6.5 |
| Oct 2024 | 8,987 | 4.58 | 1.76 | 1.0 | 9.0 | 3.3 | 4.5 | 5.7 | 7.0 |
| Jan 2025 | 4,868 | 3.51 | 1.56 | 1.0 | 9.0 | 2.4 | 3.3 | 4.4 | 5.4 |

Figure 1 Binned Scaled Score Distribution



## 4.2 Test Results by Demographic Variables

### 4.2.1 Variation by Demographic Group

Pearson VUE undertakes several tasks as part of the item development and analysis process to ensure the test content does not cause differential performance related to demographic characteristics. 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 adhere 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 test. Practice resources are also produced, and these are freely accessible to all. Finally, we analyse the performance of individual items by demographic characteristics and remove any items that might exhibit bias (as discussed in Section 6.2). The demographic information is collected via a survey when candidates register for the test. The survey questions asked, as well as the section in this report where this information is analysed, are presented in Table 5.

Table 5 Questions Asked at Registration

| Question Asked | Section |
|---|---|
| Which of the following best describes your gender? | 4.2.2 |
| Where is your area of permanent residence? | 4.2.3 |
| What is your nationality? | 4.2.3 |
| What is your ethnic group? | 4.2.4 |
| Is English your first language? | 4.2.5 |
| What is the best description of the most recent school/college you attend/attended? | 4.2.6 |
| Are you currently, or have you been, in receipt of free school meals during your secondary education? | 4.2.7 |
| Do any of your parents, step-parents or guardians have higher education qualifications, such as a degree, diploma or certificate of higher education? | 4.2.8 |
| Do you have a learning difficulty (e.g. dyslexia, dyspraxia) or any physical or mental health conditions or illnesses lasting or expected to last for 12 months or more? | 4.2.9 |

### 4.2.2 Gender

Figure 2 presents the breakdown of test-takers by gender. This shows that 70% of the candidates identified as “man” and 29% as “woman”. Male candidates slightly outperformed female candidates with a mean score of 4.31 compared to 3.93 (Table 6). This is in line with expectations based on national trends for mathematics and physics examinations.

The scaled score distribution for male and female candidates is plotted in Figure 3 and shows a slight shift to higher scaled scores for male candidates.

Figure 2 Percent of Candidates by Gender

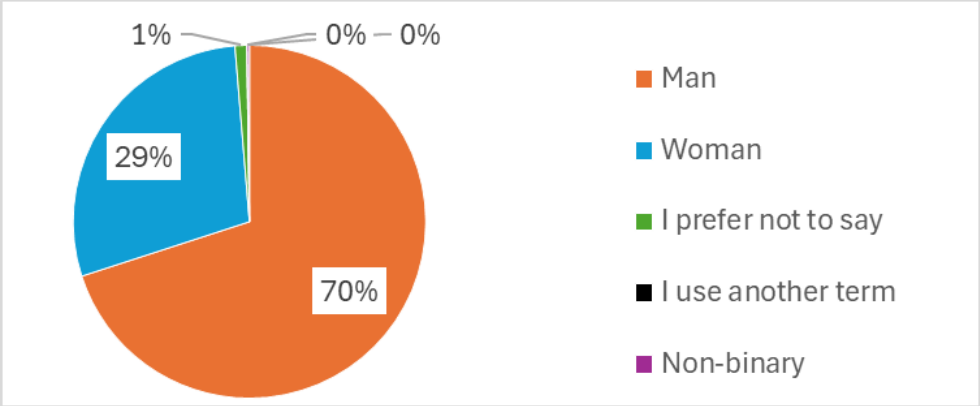
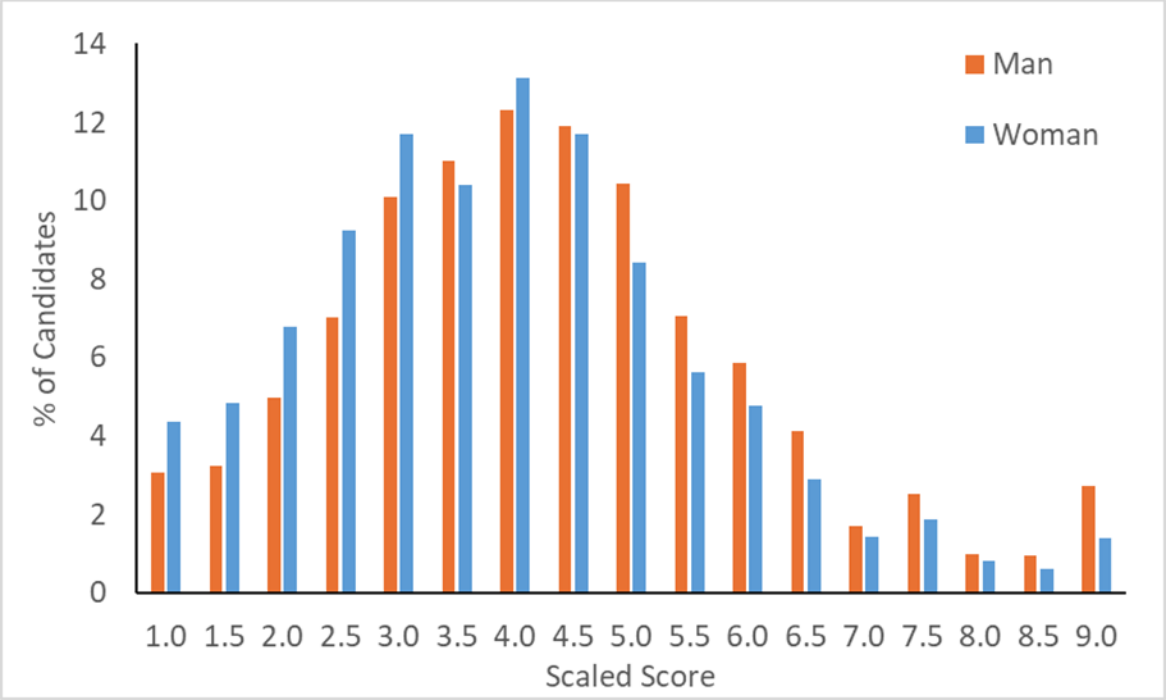



Table 6 Scaled Score Summary Statistics by Gender

| Gender | *N* | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| Man | 9,699 | 4.31 | 1.78 | 1.0 | 9.0 | 3.1 | 4.1 | 5.4 | 6.7 |
| Woman | 3,971 | 3.93 | 1.71 | 1.0 | 9.0 | 2.7 | 3.8 | 5.0 | 6.2 |

Figure 3 Scaled Score Distribution by Gender



### 4.2.3 Area of Residence

Candidates were required to state their area of residence, and these are categorised as UK, EU or Rest of World. Most candidates who took the TMUA reside in the UK, as can be seen in Figure 4 below. The scaled score summary statistics can be found in Table 7 and show that the candidates outside of the UK and the EU performed more strongly, with a mean scaled score of 4.74 compared to 3.78 for candidates in the EU and 3.86 in the UK. This can also be seen in Figure 5, where the scaled score distribution of "Other" region candidates is shifted towards higher scores.

Figure 4 Percent of Candidates by Region

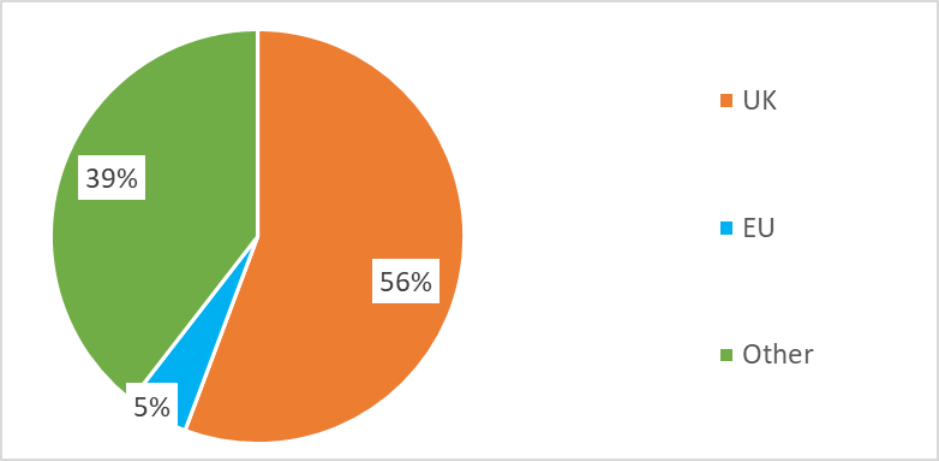
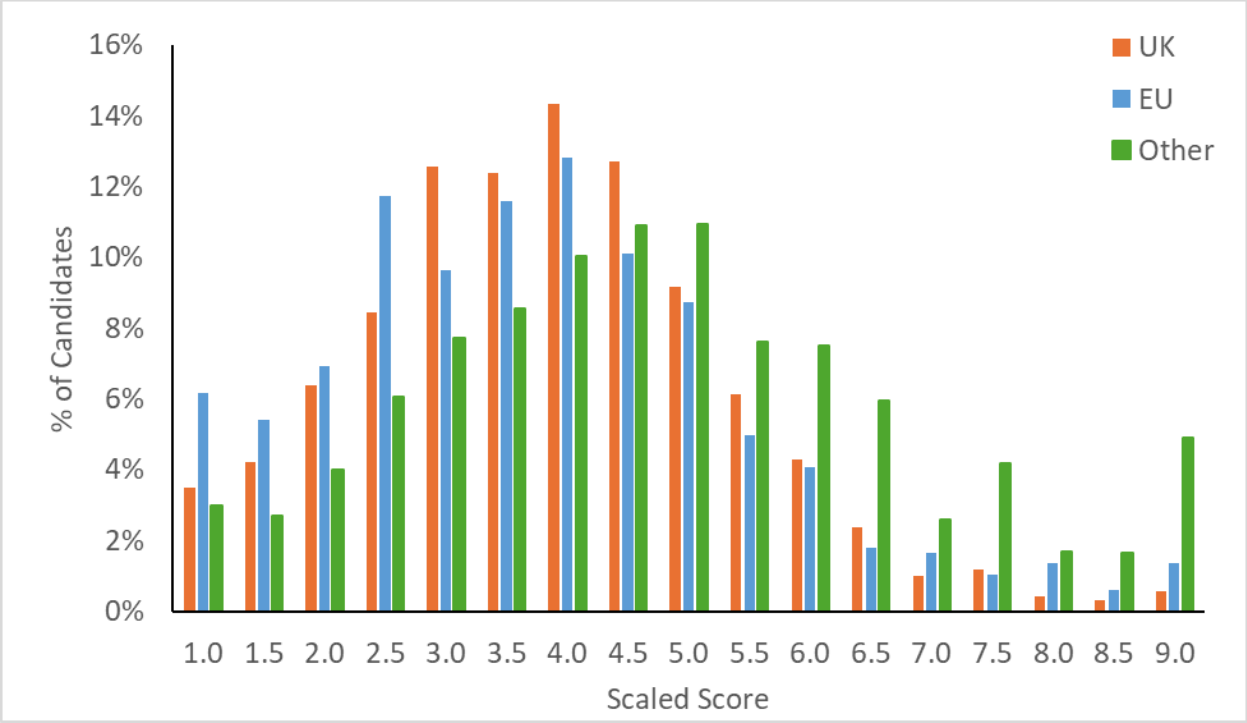



Table 7 Scaled Score Summary Statistics by Area of Residence

| Area | *N* | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| UK | 7,715 | 3.86 | 1.52 | 1.0 | 9.0 | 2.8 | 3.8 | 4.8 | 5.8 |
| EU | 664 | 3.78 | 1.73 | 1.0 | 9.0 | 2.5 | 3.6 | 4.8 | 5.9 |
| Other | 5,472 | 4.74 | 1.96 | 1.0 | 9.0 | 3.3 | 4.6 | 5.9 | 7.5 |

Figure 5 Scaled Score Distribution by Area of Residence




Candidates from outside the UK were asked to identify their nationality as well as their area of permanent residence. From all candidates, 56% identified as British, followed by 18% as Chinese and 6% as Indian (Figure 6). The summary statistics for the non-British nationalities are shown in Table 8 for the top five nationalities. This shows that candidates identifying as Chinese, Singaporean and from Hong Kong all had average scaled scores above 4.20, which is the overall average across all groups. The performance of these nationalities is likely the driving force behind the higher mean scores for candidates from outside the EU and UK.

Figure 6 Percent of Candidates by Nationality

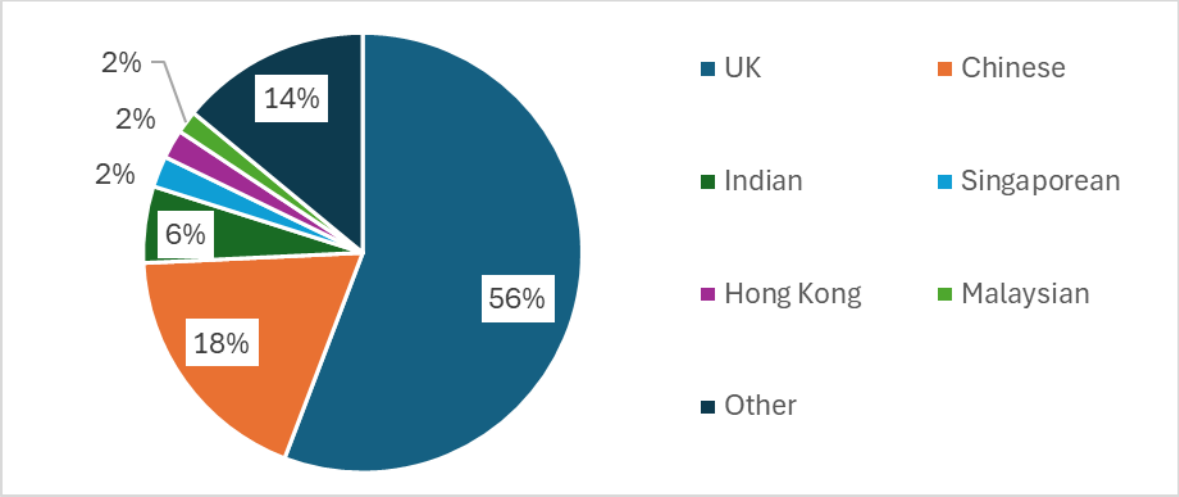


Table 8 Scaled Score Summary Statistics by Nationality

| Nationality | *N* | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| Chinese | 2,554 | 5.42 | 1.89 | 1.0 | 9.0 | 4.1 | 5.4 | 6.7 | 8.4 |
| Indian | 779 | 3.63 | 1.65 | 1.0 | 9.0 | 2.4 | 3.5 | 4.7 | 5.7 |
| Singaporean | 316 | 4.78 | 1.59 | 1.0 | 9.0 | 3.6 | 4.7 | 5.8 | 6.9 |
| Hong Kong | 296 | 5.06 | 1.81 | 1.0 | 9.0 | 3.8 | 5.0 | 6.3 | 7.6 |
| Malaysian | 231 | 3.80 | 1.53 | 1.0 | 8.8 | 2.7 | 3.8 | 4.7 | 5.7 |

### 4.2.4 Ethnicity (UK Candidates Only)

UAT-UK candidates who reside in the UK are requested to answer a question relating to their ethnicity. The categories used are:

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

Figure 7 shows the breakdown of candidates by ethnicity in the TMUA. The largest ethnic group amongst UK candidates was White (41%), closely followed by Asian (39%). Differences in mean scaled scores across ethnic groups within the UK tend to reflect underlying trends within the population. UK-White candidates had the highest mean scores, followed by UK-Asian. Figure 8 shows the cumulative percentage by scaled score for each ethnic group for UK candidates. This illustrates the range across the score distribution, with candidates identifying as UK-Black scoring generally lower than other groups and those identifying as UK-White scoring the highest. This difference is widest around the 50% mark, which is also illustrated by the median scores by ethnic group, ranging from 2.9 for UK-Black to 4.0 for UK-White. Note that these values are only for the UK candidates and could change significantly if candidates outside the UK were included given the data by nationality (4.2.3).

Figure 7 Percent of UK Candidates by Ethnicity

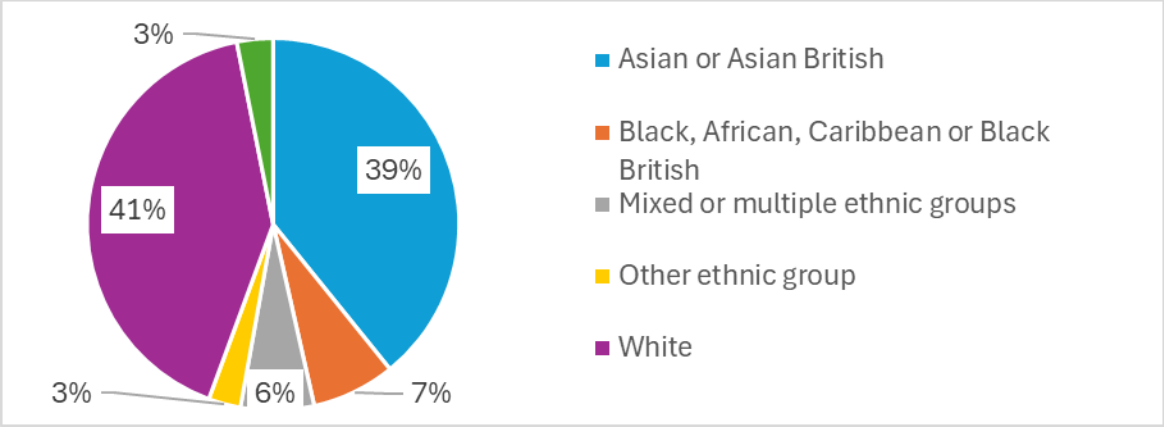
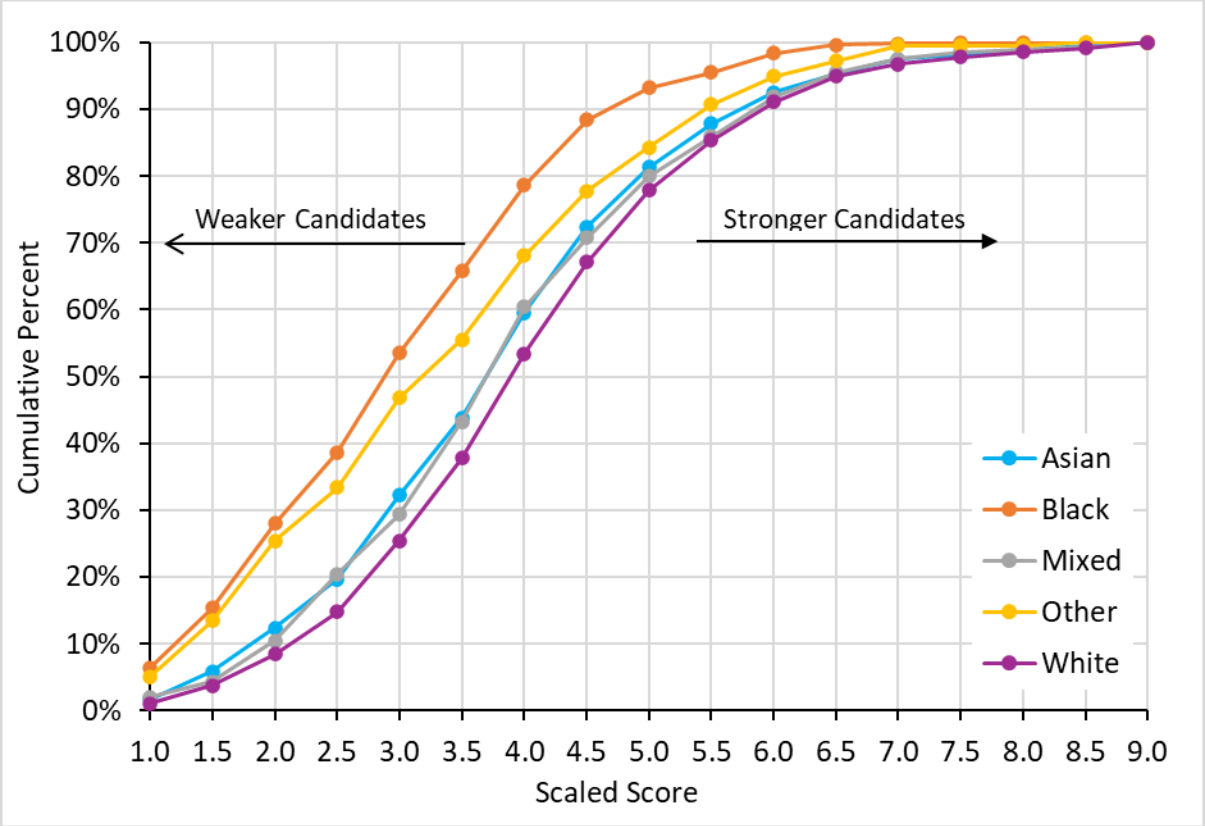


Table 9 Scaled Score Summary Statistics by Ethnicity for UK Candidates

| Ethnicity | *N* | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| Asian | 3,025 | 3.82 | 1.51 | 1.0 | 9.0 | 2.8 | 3.8 | 4.7 | 5.7 |
| Black | 561 | 3.01 | 1.31 | 1.0 | 7.4 | 2.0 | 2.9 | 3.9 | 4.7 |
| Mixed | 490 | 3.86 | 1.51 | 1.0 | 9.0 | 2.8 | 3.8 | 4.8 | 5.9 |
| Other | 216 | 3.38 | 1.56 | 1.0 | 8.4 | 2.0 | 3.3 | 4.5 | 5.5 |
| White | 3,180 | 4.04 | 1.49 | 1.0 | 9.0 | 3.0 | 4.0 | 5.0 | 5.9 |

Figure 8 Cumulative Percent Scaled Score by Ethnicity



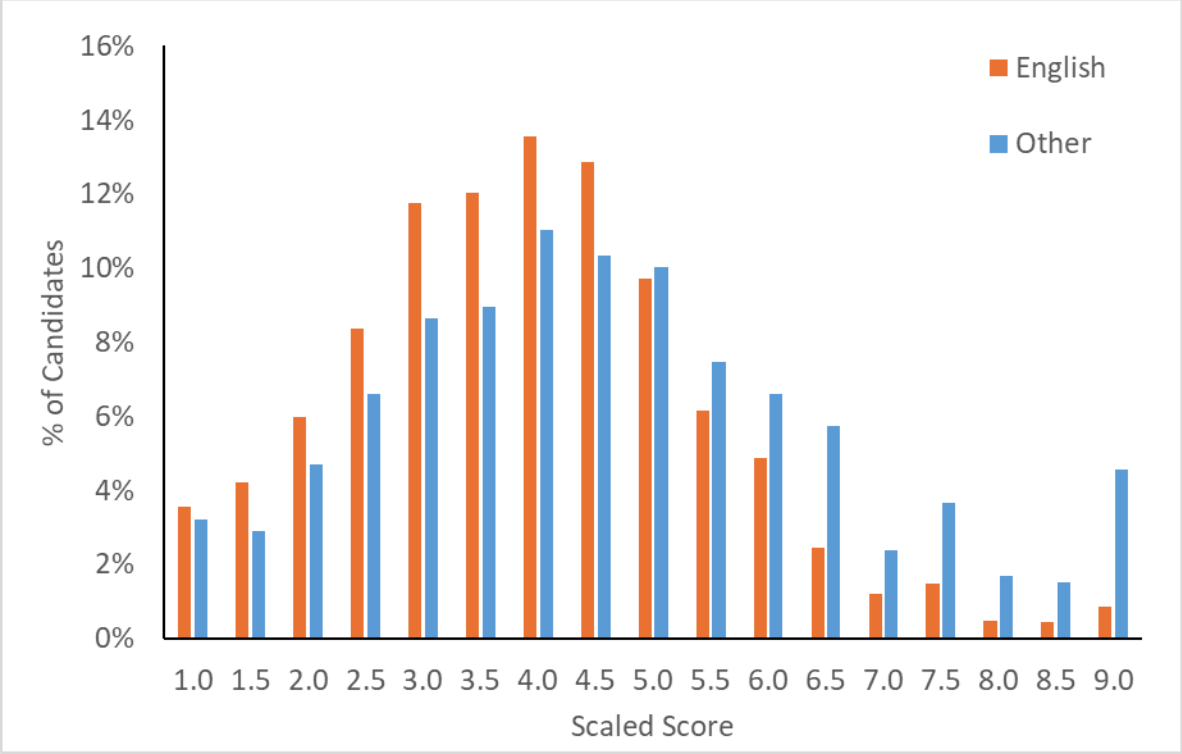
### 4.2.5 First Language

In 2024/2025, 60% of candidates identified English as their first language. Interestingly, candidates with English as a first language were weaker overall, with a mean scaled score of 3.94 compared to 4.61 for those whose first language was not English (Table 10). This difference can also be seen clearly in the scaled score distributions in Figure 9. This is an interesting finding and most likely reflects the low language load for this test, as most of the questions are very mathematical.

Table 10 Scaled Score Summary Statistics by First Language

| First Language | *N* | % | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| English | 8,373 | 60% | 3.94 | 1.58 | 1.0 | 9.0 | 2.8 | 3.8 | 5.0 | 5.9 |
| Other | 5,478 | 40% | 4.61 | 1.96 | 1.0 | 9.0 | 3.1 | 4.4 | 5.9 | 7.3 |

Figure 9 Scaled Score Distribution by First Language



### 4.2.6 Education (UK Candidates Only)

UK candidates were asked to identify their current or most recent school type. The most common school type, for 43% of candidates, was a Further Education College or Sixth Form College (Figure 10). In terms of ability, the strongest group was the 19% of candidates who attended a Private or Independent School, as this group had a mean scaled score of 4.26 (Table 11). Those attending a Further Education College or Sixth Form College had the lowest mean scaled score of 3.62 (Figure 11). This is in line with national trends, where selective or fee-paying schools tend to have stronger exam results compared to state schools, which typically educate students with a broader range of academic ability and financial backgrounds.

Figure 10 Percent of Candidates by School Type

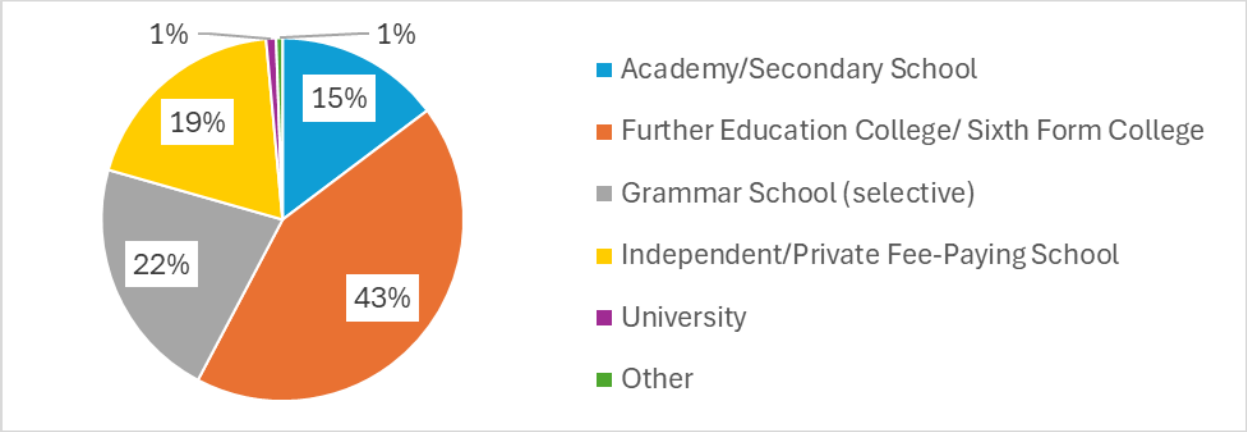
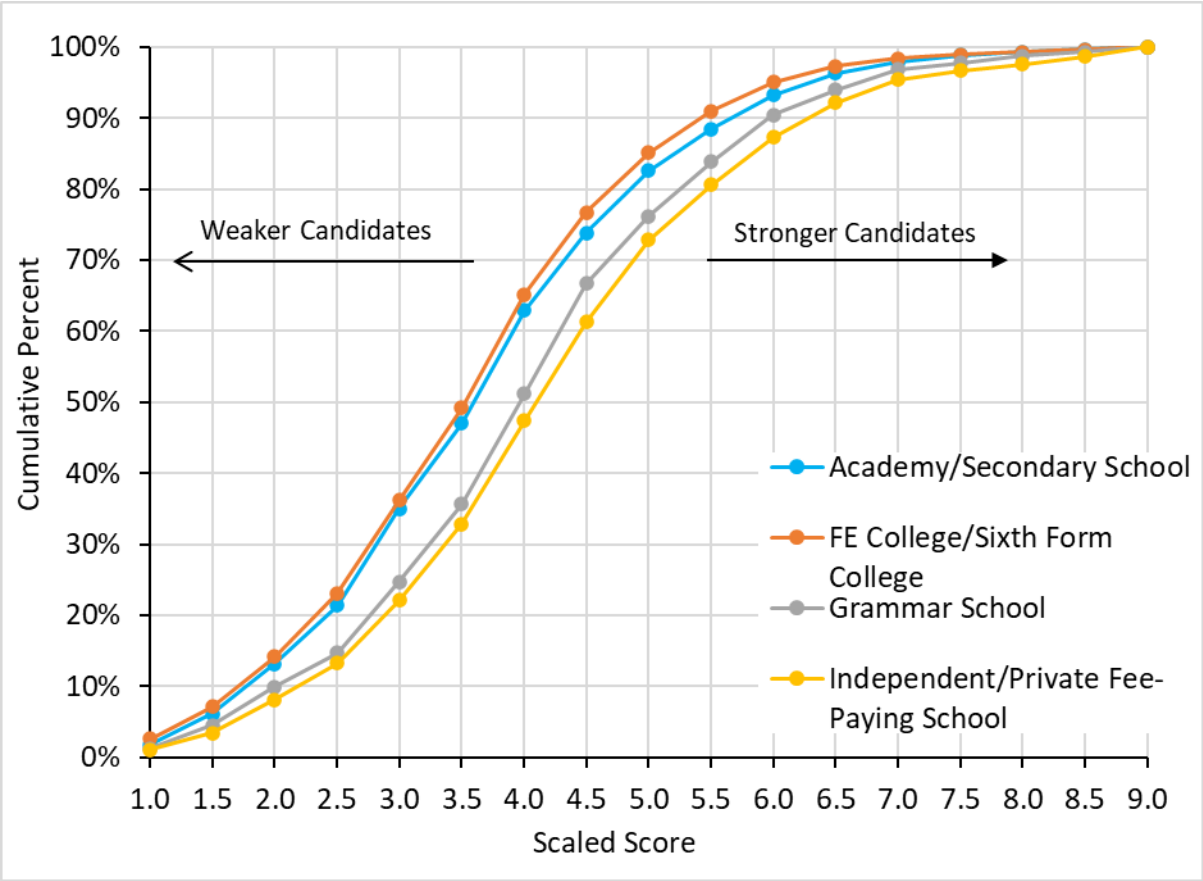



Table 11 Scaled Score Summary Statistics by School Type

| School Type | *N* | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| Secondary School | 1,138 | 3.72 | 1.48 | 1.0 | 9.0 | 2.6 | 3.6 | 4.7 | 5.7 |
| Further Education/ 6th Form College | 3,319 | 3.62 | 1.43 | 1.0 | 9.0 | 2.6 | 3.6 | 4.5 | 5.4 |
| Grammar School | 1,670 | 4.09 | 1.53 | 1.0 | 9.0 | 3.1 | 4.0 | 5.0 | 5.9 |
| Private Fee-Paying School | 1,475 | 4.26 | 1.59 | 1.0 | 9.0 | 3.1 | 4.2 | 5.2 | 6.2 |
| University | 69 | 3.71 | 1.84 | 1.0 | 8.4 | 2.2 | 3.3 | 4.7 | 6.5 |

Figure 11 Cumulative Percent Scaled Score by School Type



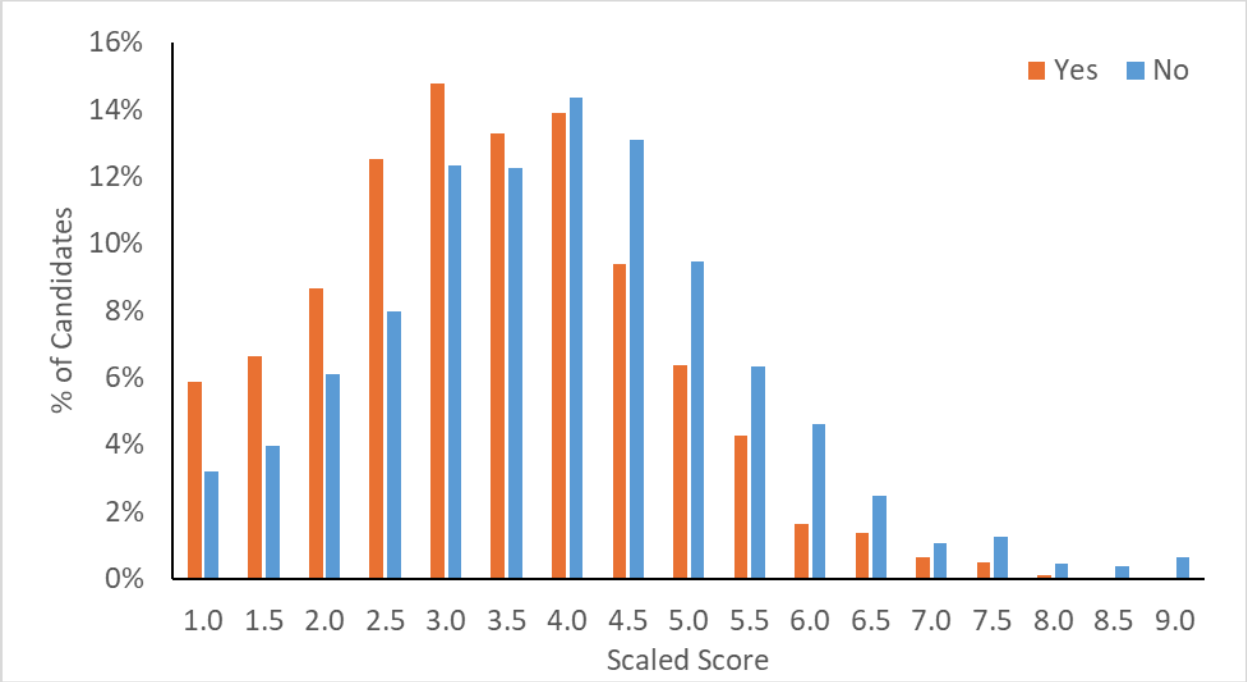
### 4.2.7 Free School Meals (UK Only)

UK candidates are asked if they are or were in receipt of free school meals during their secondary education, which is an indicator of candidates' socio-economic background. Only 10% of candidates were in receipt of free school meals, and these candidates had a lower mean scaled score than those that did not receive free school meals (Table 12). The scaled score distribution for these two groups is illustrated in Figure 12.

Table 12 Summary Statistics by Free School Meals (UK Only)

| Free School Meals | *N* | % UK | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| Yes | 798 | 10% | 3.34 | 1.36 | 1.0 | 8.1 | 2.3 | 3.3 | 4.2 | 5.2 |
| No | 6,917 | 90% | 3.92 | 1.52 | 1.0 | 9.0 | 2.8 | 3.8 | 4.9 | 5.9 |

Figure 12 Scaled Score Distribution for UK Candidates by Free School Meals (Yes/No)



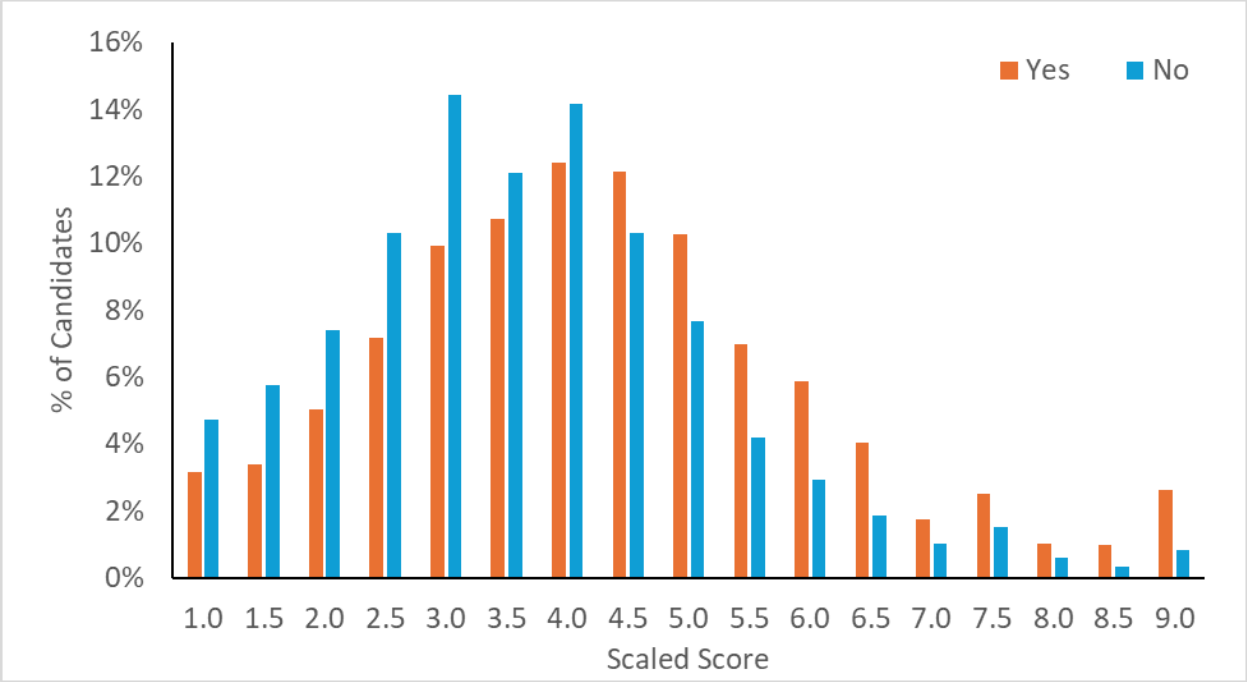
### 4.2.8 Parent Higher Education

All candidates were asked if their parent or guardian had attended tertiary education and 78% of candidates responded "Yes" (Table 13). This measure can be considered, along with free school meals, as an indicator of widening participation of candidates. As can be seen in Figure 13, the candidates who had a parent/guardian attend higher education outperformed those who did not.

Table 13 Summary Statistics by Parent/Guardian Education

| Parent Higher Education | *N* | % | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| Yes | 10,852 | 78% | 4.29 | 1.78 | 1.0 | 9.0 | 3.0 | 4.1 | 5.4 | 6.7 |
| No | 1,844 | 13% | 3.65 | 1.57 | 1.0 | 9.0 | 2.6 | 3.5 | 4.5 | 5.7 |
| Don't know | 612 | 4% | 3.98 | 1.80 | 1.0 | 9.0 | 2.6 | 3.8 | 5.2 | 6.4 |
| I prefer not to say | 543 | 4% | 4.51 | 1.73 | 1.0 | 9.0 | 3.3 | 4.5 | 5.7 | 6.7 |

Figure 13 Scaled Score Distribution by Parent/Guardian Higher Education



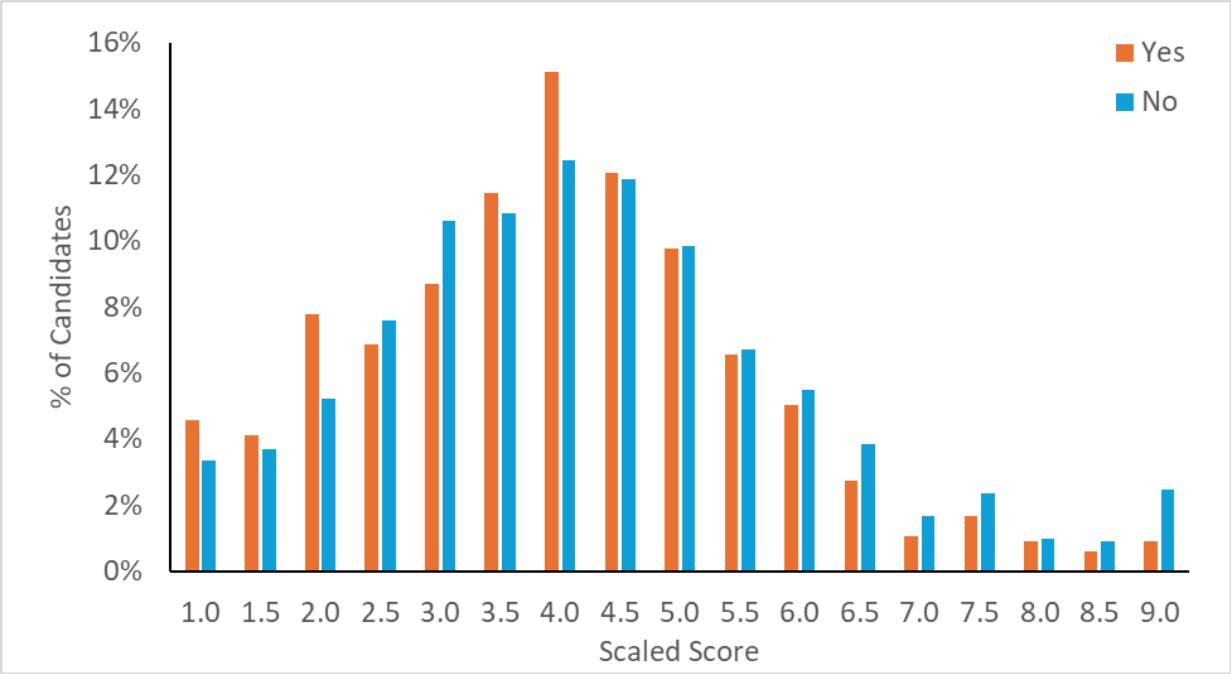
### 4.2.9 Learning Difficulty/Chronic Heath Condition

Candidates are asked whether they have a learning difficulty or health condition that has lasted (or is expected to last) for a year or longer. The majority of candidates, 91%, said that they did not have such a learning difficulty or health condition (Table 14). There is a difference in performance across the two groups with those who responded "No" outperforming those who responded "Yes" (Figure 14).

Table 14 Summary Statistics by Learning Difficulty/Chronic Health Condition

| Learning Difficulty/Health Condition | *N* | % | Scaled Score | | | | Percentile | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Min | Max | 25 | 50 | 75 | 90 |
| Yes | 655 | 5% | 3.97 | 1.65 | 1.0 | 9.0 | 2.8 | 3.9 | 5.0 | 6.1 |
| No | 12,549 | 91% | 4.22 | 1.78 | 1.0 | 9.0 | 2.9 | 4.0 | 5.2 | 6.6 |
| Don't know | 323 | 2% | 3.78 | 1.61 | 1.0 | 9.0 | 2.6 | 3.6 | 4.9 | 5.9 |
| I prefer not to say | 324 | 2% | 4.42 | 1.78 | 1.0 | 9.0 | 3.0 | 4.3 | 5.7 | 7.0 |

Figure 14 Scaled Score Distribution by Candidates with a Learning Difficulty or Chronic Health Condition



## 4.3 Accommodations

Candidates are able to request accommodations, such as extra time, if required. In total, 517 candidates had accommodations for TMUA, which is around 4% of the candidate population. Candidates can apply for more than one accommodation. Table 15 summarises the number of accommodations by the type required, but note that as candidates can apply for more than one accommodation, some are counted more than once. The most common request was for extra time and/or pause-the-clock, with 512 candidates (99% of accommodations candidates) requesting this.

Table 15 T Number of Accommodations by Type

| Accommodation | *N* |
|---|---|
| Extra Time | 337 |
| Pause-the-clock | 55 |
| Extra time + pause-the-clock | 120 |
| Separate room | 118 |
| Other | 87 |

# 5. Test Level Analysis

## 5.1 Reliability and *SEM*

Reliability, as it applies to testing, can be thought of as the consistency, or reproducibility, of test scores. A common estimate of test score reliability is Cronbach's alpha ($\alpha$), which is an indicator of the test's internal consistency. Cronbach's alpha, which ranges from 0 to 1, is based on the degree of score intercorrelation among the items on the test. A higher $\alpha$ suggests that similar results would probably be observed if a given candidate was administered the same (or an equivalent) test form on a different occasion. A general rule of thumb is that $\alpha$ should be at least 0.80 (Nunnally & Bernstein, 1994). However, this is also dependent on the length of the test as reliability tends to increase as test length increases.

The *SEM* allows us to create a confidence band for the candidate's hypothetical 'true' score, which is defined as the average of a candidate's scores if he or she were to take the same (or a parallel) test many times. In general, the smaller the *SEM* for the test, the more confidence one can place in the assigned scores. The *SEM* is calculated via the following equation:

$$SEM = \sigma_x \sqrt{1 - r_{xx}}$$

where $\sigma_x$ is the standard deviation (*SD*) of the raw (number-correct) scores and $r_{xx}$ is the reliability estimate.

Under classical measurement theory, there is approximately a 95% probability that a candidate's true score lies within +/- 2 *SEM*s of his or her observed score on a particular test administration, and approximately a 68% probability that it lies within +/- 1 *SEM*. The raw score reliability, or Cronbach's alpha, and the raw score *SEM* for each form can be found in Table 16.

The raw score reliabilities are excellent, ranging from 0.77 to 0.87, with all but one form being above 0.80. The alpha of 0.77 could be due to a lower mean point biserial on the items on this form, which could be improved by selecting better discriminating items.

Table 16 Raw Score Reliability and SEM

| Event | Raw Score Reliability | |
|---|---|---|
| | Cronbach's Alpha | *SEM* |
| Oct 2024 | 0.81 to 0.87 | 2.61 to 2.72 |
| Jan 2025 | 0.77 to 0.84 | 2.59 to 2.67 |

As TMUA candidates also receive a scaled score, which is scaled from the candidate theta, the reliability of the theta estimate is also important when assessing scaled scores. The scaled score reliability and SEM are summarised in Table 17. This shows that the scaled score reliabilities are very similar to the raw score reliabilities. They range from 0.76 to 0.87, with all but one form having a reliability over 0.80.

Table 17 Scaled Score Reliability and SEM

| Event | Scaled Score Reliability | |
|---|---|---|
| | **Reliability** | ***SEM*** |
| Oct 2024 | 0.82 to 0.87 | 0.67 to 0.68 |
| Jan 2025 | 0.76 to 0.85 | 0.68 to 0.70 |

## 5.2 Test Timing Analysis

The section time for each candidate is calculated by summing the item and review time for each item and candidate for the items in the test (that is, not the non-disclosure agreement or survey). The time limit for the standard TMUA is 02:30:00. The test time summary statistics are shown in Figure 15.

A total of 90% of all candidates used between 02:15:00 and 02:30:04 for the test. The 3% of candidates with a test time exceeding the standard time are assumed to have been given a time accommodation. These candidates were excluded from the analysis of test time shown in Table 18. Test Time Summary Statistics.

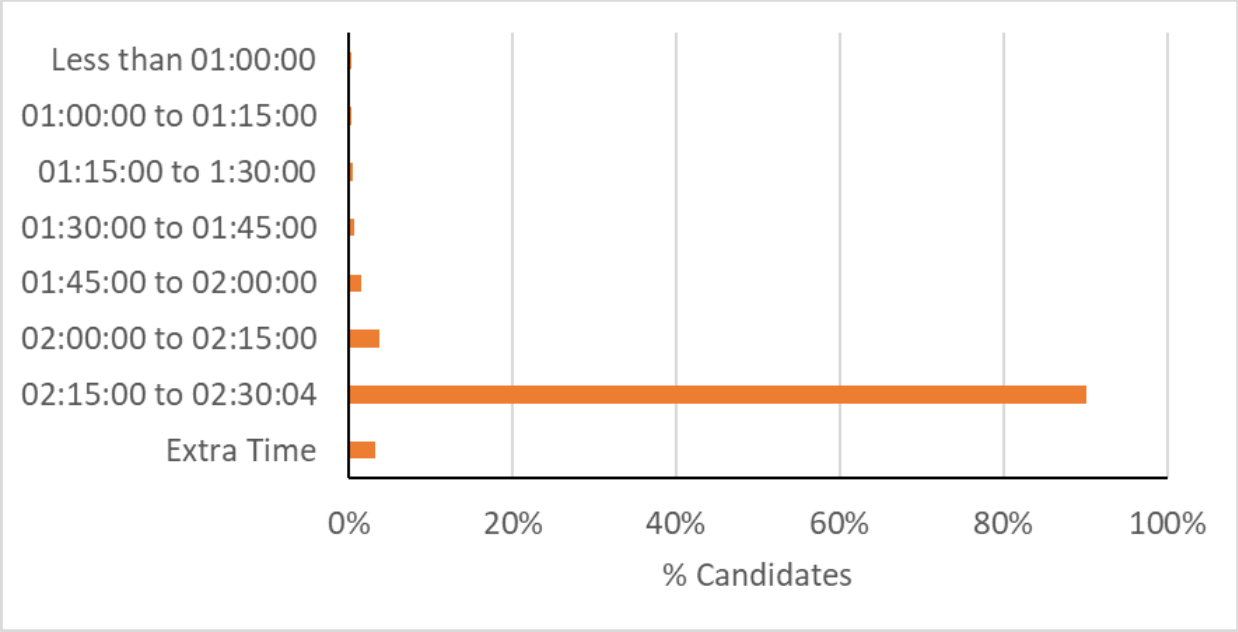
The mean test time for the 13,409 candidates included in the test time analysis was 02:26:21. However, it should be noted that there were a small number of very low test times, and this can have a significant impact on the mean time. Therefore, the median can give a better indication of the average test time. For the candidates included in the analysis of test time, the median test time is 02:29:37, which is very close to the maximum test time.

Table 18. Test Time Summary Statistics

| ***N*** | **Test Time** | | | | |
|---|---|---|---|---|---|
| | **Mean** | **Median** | ***SD*** | **Min** | **Max** |
| 13,409 | 02:26:21 | 02:29:37 | 00:10:49 | 00:19:39 | 02:30:03 |

*Note: 446 candidates were excluded from this analysis as they had a test time over 02:30:04 and are assumed to have had a time accommodation*

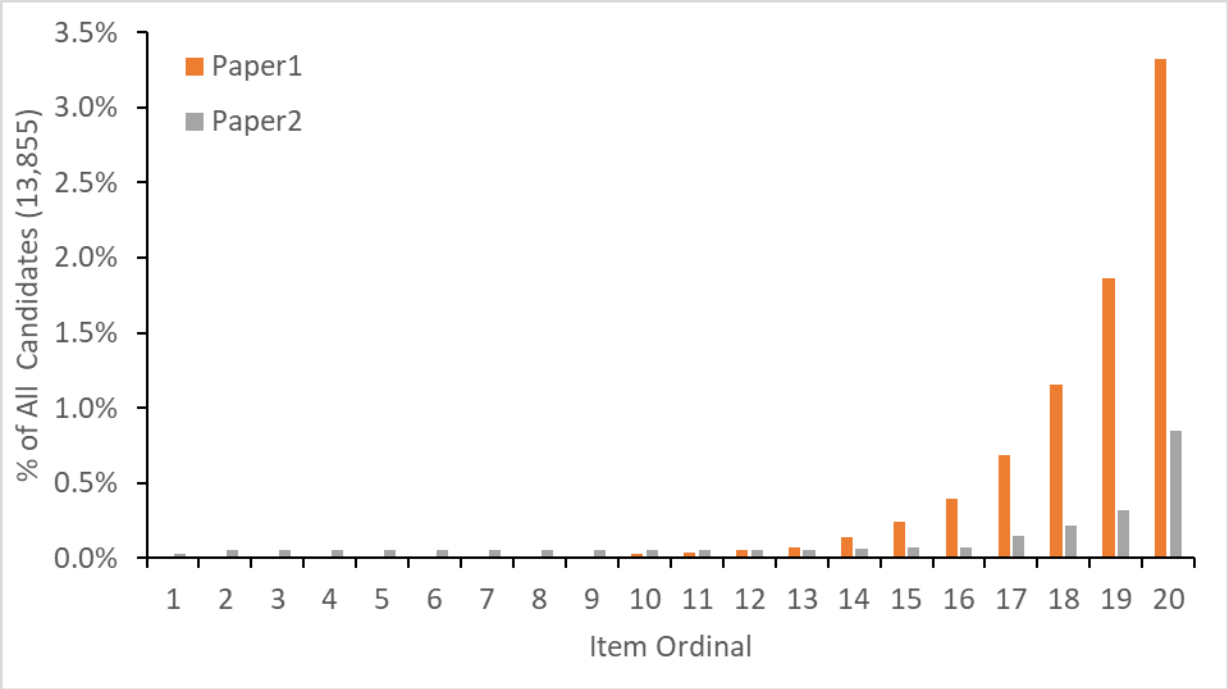
Figure 15 Percentage of Candidates by Test Time



The test timing analysis implies that the majority of candidates are using the full test time available. This indicates that the test might be speeded. Speededness can also be assessed by looking at the number of unreached items. For TMUA, the test is split into two timed sections: Paper 1 and Paper 2. Each section contains 20 items, and candidates are allotted 75 minutes per section. Since TMUA is used for selection, its aim is to help universities distinguish between highly capable candidates. Consequently, the test is expected to be speeded for some candidates, as the items are designed to progressively increase in difficulty throughout each section.

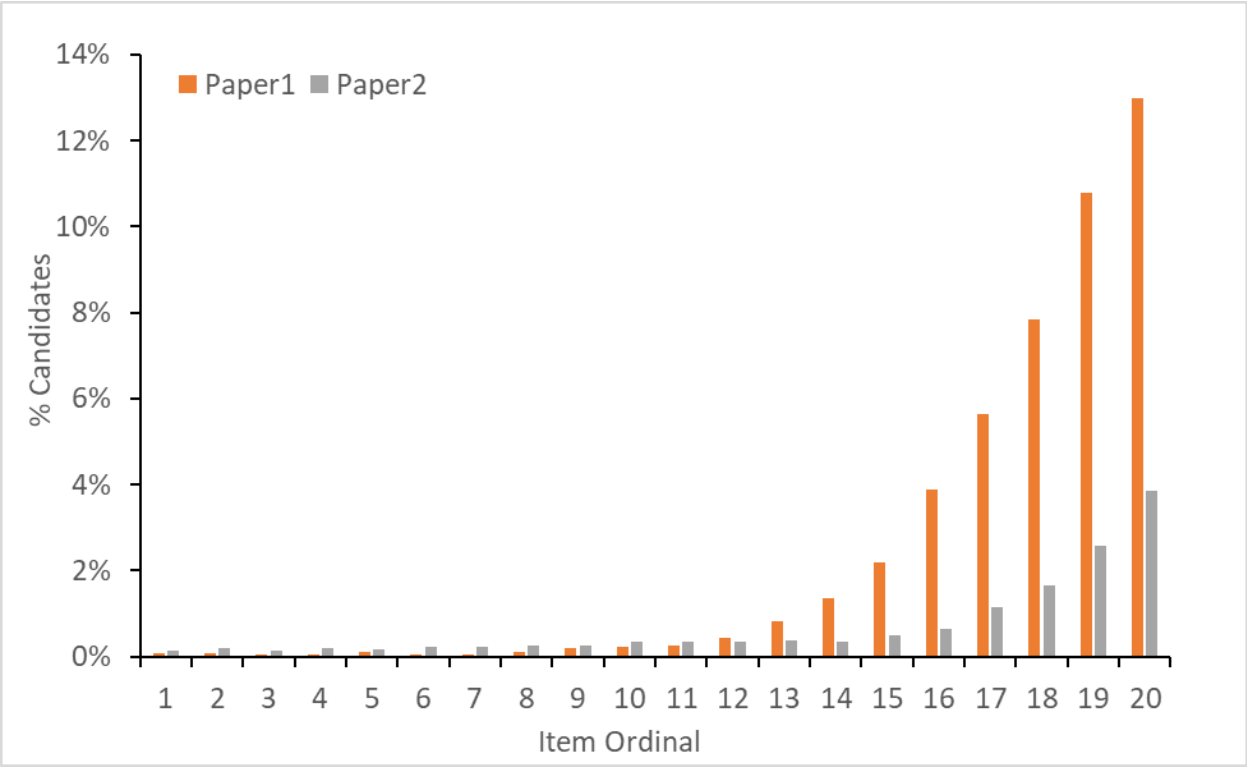
Figure 16 illustrates the number of unreached items (items not presented to the candidate because they ran out of time or ended the test) by item ordinal (sequence in section). It should be noted that the test has non-linear navigation, so candidates do not necessarily move through the test from item 1 to item 2 and so on—they can move about as they choose. A total of 468 candidates had unreached items on Paper 1 and 119 candidates had unreached items on Paper 2. This is a very small proportion of the 13,855 candidates who took the test.

Figure 16 Percent of Unreached Items by Item Ordinal by Section




When candidates are running out of time, or if they do not understand the questions, they may click through questions and randomly guess. This information would not be captured in the unreached item analysis. Therefore, item response time was also analysed to assess speededness. It is likely that candidates answering a question in less than 10 seconds were randomly guessing, so any item with a response time below or equal to 10 seconds was considered to be an unreached item (Figure 17). A total of 3,227 candidates (23%) had at least one item response time of 10 seconds or less (or item not presented). On average, candidates responded in 10 or less seconds to 2.62 items.

Figure 17 Percent of Low Time Responses (10 seconds or less) by Item Ordinal in Section



# 6. Item Performance

Each year, Pearson VUE undertakes item writing, data analysis and statistical screening. At the end of each testing window, all items are analysed. The purpose of item analysis is to examine the item quality.

## 6.1 TMUA Item Analysis

For TMUA, item quality is assessed on three statistical criteria:

- Point biserial: the degree to which a test item discriminated between strong and weak candidates. Point biserial ranges from -1 to +1, with positive values indicating that the item discriminates well.
- *p* Value: the proportion of candidates who answered the item correctly—the item difficulty. This index of difficulty is dependent on the candidate population that saw the item and can therefore be influenced by the candidate sample. Ideally items should have a value between 0.20 and 0.90, although a wider range would be acceptable on this test as the purpose is to stretch the scale and there are some very high scoring candidates.
- IRT *b*: the difficulty parameter from the IRT analysis of the items. Ideally items should have a *b* value between -3 and +3.

Items that do not meet the statistical criteria above are subjected to further scrutiny to determine if the key (stated correct answer) is correct, or if the item is flawed in some way. Such items are typically then used for training purposes to show item writers what type of item does not work well.

Of the items used across the TMUA forms, 99% met the criteria (Table 19). This is well above the 80% typical target for new exams.

Table 19 Item Status Outcome

| Status | Comment | *N* Items | % of Items |
|---|---|---|---|
| Fail | Very low point biserial | 1 | 1% |
| Pass | *p* Value greater than 0.90 | 1 | 1% |
| | *p* Value less than 0.20 | 21 | 11% |
| | None | 163 | 88% |

## 6.2 Differential Item Functioning (DIF)

### 6.2.1 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 candidate ability—possibly some characteristic of the candidates that is related to gender.

The UAT-UK DIF comparison groups are based on:

- Gender: Male vs Female
- UK Ethnicity: White vs non-White
- UK School Type: Academy/Further Education College vs Grammar/Private School
- First Language: English vs non-English

For the analysis by UK ethnicity and UK school type, several groups were combined to provide a sufficient volume for the analysis. For UK ethnicity, the non-White group will be dominated by UK-Asian as this is the next largest group. The grouping by UK school type was determined by earlier analysis, as the performances of candidates at an Academy or a Further Education College were similar to each other as were Grammar and Private School candidates.

The remaining demographic categories did not have sufficient numbers of candidates for analysis.

### 6.2.2 Method of DIF Detection

For TMUA, the Mantel–Haenszel (MH) 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.

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 this test, 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 identified as Category C are flagged for review as they may contain bias. Items in Categories A and B are not flagged because of the small effect or lack of statistical significance.

### 6.2.3 Sample Size Requirements

The minimum sample size requirements used for the UAT-UK DIF analyses were at least 50 candidate responses per group and at least 200 responses in total. If the sample size for the DIF analysis is less than 200, the sample is not large enough to analyse and therefore DIF is not reported.

### 6.2.4 DIF Results

The DIF results are reported in Table 20 for items that showed Category C DIF. These are items where there is a significant difference in the performance of candidates in different demographic groups. One item was identified as showing significant DIF by gender, with candidates who identified as male outperforming female candidates, after controlling for differences in ability. This was a very difficult item—one of the hardest on the form—but this analysis compares candidates with an overall similar ability, so the difficulty of the item should not have an impact. For this item, almost all candidates responded (only 9 skipped it), and for candidates who identified as male the *p* value was 0.19 compared to 0.07 for female candidates. This item will be reviewed to identify likely sources of bias, and this information used to inform future item writing.

Eleven items were flagged as showing DIF by first language, with five favouring non- English candidates and six favouring candidates with English as their first language. These items are across a range of item difficulties. These items will be reviewed for possible bias to inform future item writing.

No significant DIF was identified by UK School Type or UK Ethnicity.

Table 20 Items Flagged with Category C DIF

| Category | Item | Item *b* Value | Preferred Group | MH DIF Value (> 1.5 Category C) | *p* Value (significance < 0.001) |
|---|---|---|---|---|---|
| Gender | 733927 | 1.9899 | Male | -1.83300 | 0.0006 |
| English as a First Language | 733548 | -0.4910 | Not English | 1.72020 | 0.0000 |
| | 733781 | 0.4937 | Not English | 1.62150 | 0.0000 |
| | 733829 | 0.5011 | English | -1.60740 | 0.0004 |
| | 733839 | 0.1441 | Not English | 1.91055 | 0.0001 |
| | 733864 | 0.1230 | Not English | 1.73900 | 0.0000 |
| | 733870 | 0.7295 | Not English | 2.00455 | 0.0000 |
| | 733919 | -0.7953 | English | -1.65205 | 0.0000 |
| | 733948 | -0.7993 | English | -1.64970 | 0.0002 |
| | 733970 | 0.1064 | English | -1.93405 | 0.0000 |
| | 734144 | 0.0028 | English | -1.52750 | 0.0000 |
| | 739277 | 1.0889 | English | -1.85650 | 0.0000 |

## 7. References

Linacre, J. M. (2014). Winsteps Rasch measurement computer program. Beaverton, OR: Winsteps.com.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York, NY: McGraw-Hill.