University Admissions Tests UK

UAT-UK Ltd

Annual Reports 2025-2026

TARA Technical Report

Published July 2026

Contents

1. Executive Summary

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

The Test of Academic Reasoning for Admissions (TARA) was administered in two sittings: 15th and 16th October 2025, and 12th and 13th January 2026, for candidates applying to start university in 2026. A total of 2,626 candidates completed the assessment, which comprises three modules: Critical Thinking, Problem Solving, and an unscored Writing Task. Scaled scores from 1.0 to 9.0 were awarded for the two scored modules, with scaling anchored to the October candidate ability distribution to ensure comparability across sittings within the cycle. Candidate performance across both sittings showed normally distributed scaled scores for each module, enabling effective differentiation among applicants.

When candidates register for the test, they complete a questionnaire about their demographic characteristics such as gender, ethnicity and school type. The analysis of these demographic variables revealed patterns consistent with national educational trends. Male candidates outperformed female candidates in both modules, with a particularly pronounced gap in the Problem Solving module. Candidates from higher socioeconomic backgrounds—measured through school type, parental education, and free school meal eligibility—achieved higher scores on average. UK candidates identifying English as their first language outperformed non-native speakers in the Critical Thinking module, reflecting the module's higher linguistic load. This difference was much less marked in Problem Solving.

Multiple forms (versions) of the TARA modules were used, and these were administered at different times to different regions. Therefore, the candidate population, and thus the scaled score summary statistics, are variable across the forms. This is expected. The TARA modules are relatively short, with 22 items and, therefore, the reliability would be expected to be lower than for a longer test. Given this, the reliabilities for the TARA modules were satisfactory, ranging from 0.59 to 0.75. The forms were well balanced in terms of difficulty, despite no item statistics being available when the items were selected for the forms.

The candidates taking TARA include very able students, so the test is designed to challenge even the best candidates with difficult questions. Given this, it is expected that the test will be slightly speeded for the majority of candidates. Test timing analysis showed that most candidates used nearly the full 40 minutes for both scored modules. Problem Solving was more speeded than Critical Thinking, with 8% of candidates leaving at least one item unreached compared with 1% in Critical Thinking. Low time responses (under 5 seconds) followed a similar pattern.

Item performance was strong across both modules. All items met the criteria — well above the 80% target for new items. Items demonstrated good discrimination, with high point biserial values and an appropriate spread of difficulty. Differential item functioning (DIF) analysis identified no Critical Thinking items with significant DIF, but one Problem Solving item showed gender related DIF, which will be reviewed.

Overall, the 2025/26 TARA administration was successful. The test demonstrated sound psychometric properties, effective differentiation across the ability range, and strong item performance.

2. Introduction

The Test of Academic Reasoning for Admissions (TARA) is a non-subject-specific test used across a range of subject areas. It is used by universities to distinguish between a large number of strong candidates with similar academic profiles.

The TARA is available in two sittings within each admissions cycle, and this report covers those candidates who took the exam in October 2025 (15th and 16th) and January 2026 (12th and 13th). The test consists of three modules: Critical Thinking, Problem Solving, and the Writing Task. All candidates are required to take all modules. For the Critical Thinking and Problem Solving modules, candidates receive a scaled score from 1.0 to 9.0 with no overall score. The Writing Task is unscored and is passed on to universities to use as they wish. Section 3 outlines the structure of the test and the measurement approach used for it.

Section 4 describes the test results, including the overall scaled score results, the proportion of candidates requiring access arrangements, 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) at module level, 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 the results of a differential item functioning (DIF) analysis by demographic variables where there were sufficient candidates.

3. Test Design and Measurement Approach

3.1 Test Design

The 2025/2026 TARA contained three separate modules as summarised in Table 1. A number of forms, or versions, of each module are available during the testing 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 TARA Test Design

Test	Module	Questions	Duration
TARA	Critical Thinking	22 multiple-choice questions	Each module is 40 minutes
	Problem Solving	22 multiple-choice questions	
	Writing Task	One prompt selected from a choice of three. Word limit of 750 words	

Candidates are given 40 minutes per module to answer 22 items for each of the Critical Thinking and Problem Solving modules and one item for the Writing Task. Candidates are also able to apply for extra time access arrangements if required.

Candidates are awarded a scaled score from 1.0 to 9.0, reported to one decimal place, for each module (with the exception of the Writing Task). Unlike a raw score, which is a function of 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 in Critical Thinking, for example, in either January or October has a higher ability than a candidate who scored 4.2 in Critical Thinking in either event. The scaled scores are not comparable across modules, and there is no aggregate or total score. Further details on the scoring process are provided in the subsequent sections.

3.2 Measurement Model

3.2.1 Item Analysis

The two scored TARA modules are all treated as separate tests for the analysis. For the 2025/2026 cycle, 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 each scored TARA module, 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 θ (theta) and item difficulty (also called the *b* value) by the model parameter b. Both θ 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 module 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 rank the candidates, which is the prime objective of an admissions test. For each scored TARA module, candidates are issued a scaled score ranging from 1.0 to 9.0 reported to one decimal place.

The two scored TARA modules are 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 for the October administration, and therefore the scaling is different for each admissions cycle. This means that the scaled scores from this cycle cannot be directly compared to those from future events. For UAT-UK tests, the median candidate theta is fixed to a scaled score of 4.5, and the candidate ability corresponding to the 90th percentile is fixed to a scaled score of 7.0 (Table 2). A regression line is then plotted between these two points to determine the scaling constants (Table 3) used to transform the theta values to scaled scores, which are capped at 1.0 and 9.0. The same scaling constants were used for both the October 2025 and January 2026 events to ensure the scaling was consistent and scaled scores were comparable across the two events within the 2026 admissions cycle.

Table 2 Ability Estimates Used to Scale the Scored TARA Modules

Event	Module	Percentile	Ability	Scaled Score
Oct 2025	Critical Thinking	50	1.2529	4.5
		90	2.5610	7.0
	Problem Solving	50	0.5029	4.5
		90	1.8528	7.0

Table 3 Scaling Constants

Module	Constant	Multiplier
Critical Thinking	2.1055	1.9112
Problem Solving	3.5686	1.8520

4. Test Results

4.1 Candidate Performance

This report covers test results for the 2026 admissions cycle, which includes the October 2025 event and the January 2026 event.

There were 2,626 TARA candidates in total, with 976 (37%) sitting in October 2025 and 1,650 (63%) in January 2026 (Table 4). Candidates are only allowed to sit the exam once within each admissions cycle, and, as this test was not used by the University of Cambridge, all candidates could choose to sit in either October or January. The scaled score statistics for the complete cohort and each event are summarised by module in Table 4. The two individual modules are scaled separately, and therefore, the scaled scores are not directly comparable and cannot be aggregated. In addition, as the scaling is calculated independently for each admissions cycle, the scaled score summaries cannot be directly compared across admissions cycles.

The scaled score distribution is illustrated in Figure 1 and Figure 2 for each of the modules. Candidate scaled scores are normally distributed across the scaled score range, enabling universities to effectively differentiate between the candidates.

Table 4 Scaled Score Summary Statistics

Module	Event	*N*	Mean	*SD*	Min	Max	Percentile			
							25	50	75	90
Critical Thinking	All	2,626	4.21	1.92	1.0	9.0	2.8	4.0	5.4	6.7
	Oct 2025	976	4.46	2.06	1.0	9.0	2.9	4.5	5.8	7.0
	Jan 2026	1,650	4.06	1.82	1.0	9.0	2.8	4.0	5.2	6.6
Problem Solving	All	2,626	4.72	1.77	1.0	9.0	3.5	4.7	5.8	7.0
	Oct 2025	976	4.66	1.81	1.0	9.0	3.5	4.5	5.8	7.0
	Jan 2026	1,650	4.76	1.74	1.0	9.0	3.5	4.8	5.7	7.0

Figure 1. Critical Thinking Binned Scaled Score Distribution

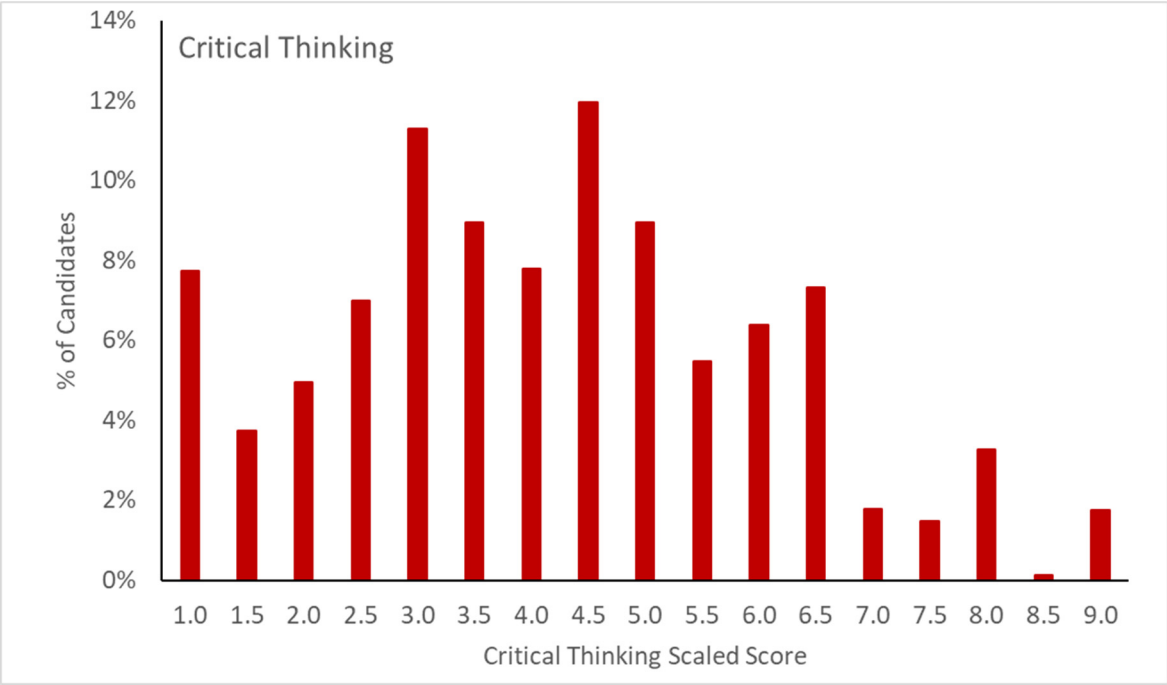
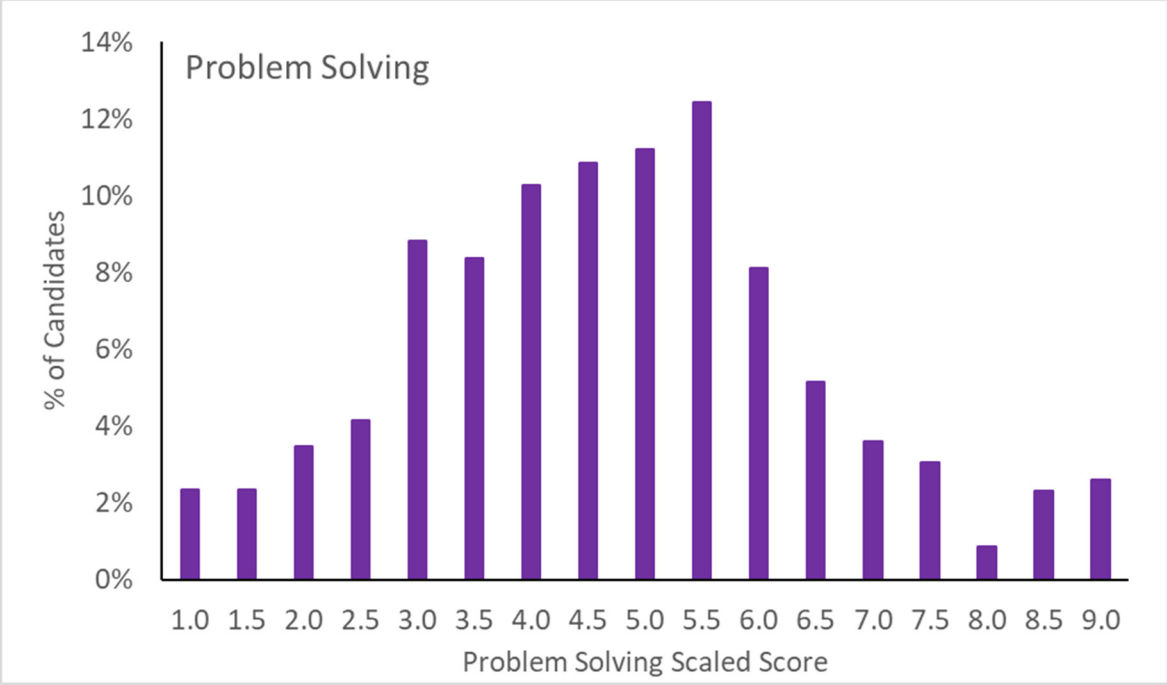


Figure 2. Problem Solving Binned Scaled Score Distribution



4.2 Test Results by Demographic Variables

4.2.1 Variation by Demographic Group

UAT-UK and Pearson undertake 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 to identify 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 country of permanent residence?	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 in secondary education and eligible for free school meals?	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 3 presents a breakdown of candidates by gender across the two modules, which shows that 62% of candidates identified as "Man" and 37% as "Woman". Male candidates outperformed female candidates across both modules and gained higher mean and median scaled scores (Table 6). The difference in mean scaled scores was 0.27 for Critical Thinking and 0.94 for Problem Solving.

Figure 3 Percentage of Candidates by Gender

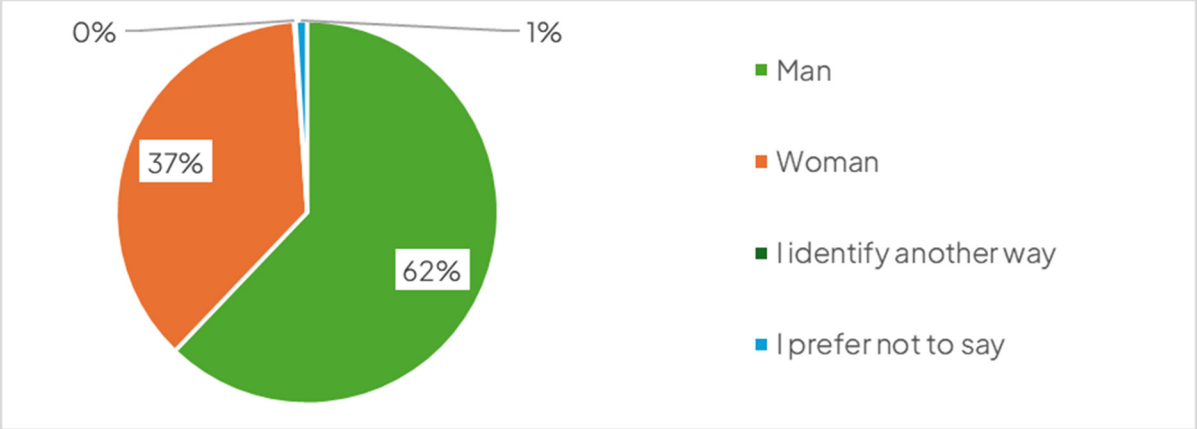
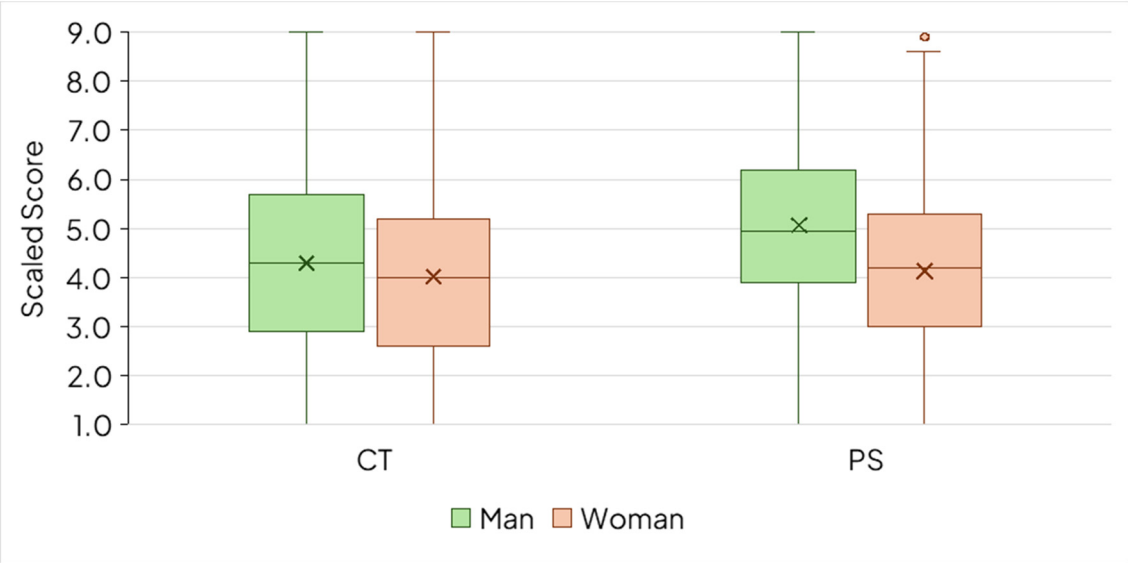


Table 6 Scaled Score Summary Statistics by Gender

Module	Gender	*N*	*%N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Critical Thinking	Man	1,632	62%	4.29	1.94	1.0	9.0	2.9	4.3	5.7	6.7
	Woman	965	37%	4.02	1.89	1.0	9.0	2.6	4	5.2	6.7
Problem Solving	Man	1,632	62%	5.06	1.77	1.0	9.0	3.9	5.0	6.2	7.6
	Woman	965	37%	4.13	1.60	1.0	9.0	3.0	4.2	5.3	6.0

Figure 4 is a box and whisker plot of scaled scores by gender for each module. The "box" shows the range of the upper and lower quartiles of the distribution (that is, the middle 50% of the data), and the "whiskers" show the minimum and maximum in-range values (excluding outliers). The cross in the box illustrates the mean scaled score, and the line illustrates the median. This plot shows that male candidates outperformed female candidates across both modules, but the gap is much more marked for Problem Solving than it is for Critical Thinking.

Figure 4 Box and Whisker Plot of Scaled Score by Gender



4.2.3 Area of Residence

Candidates were required to state their area of residence, and these have been categorised as UK, EU, or Other (Rest of World). TARA candidates were fairly evenly split between the UK (45%) and Other (46%) with only 9% from the EU (Figure 5).

Figure 5 Percentage of Candidates by Area of Residence

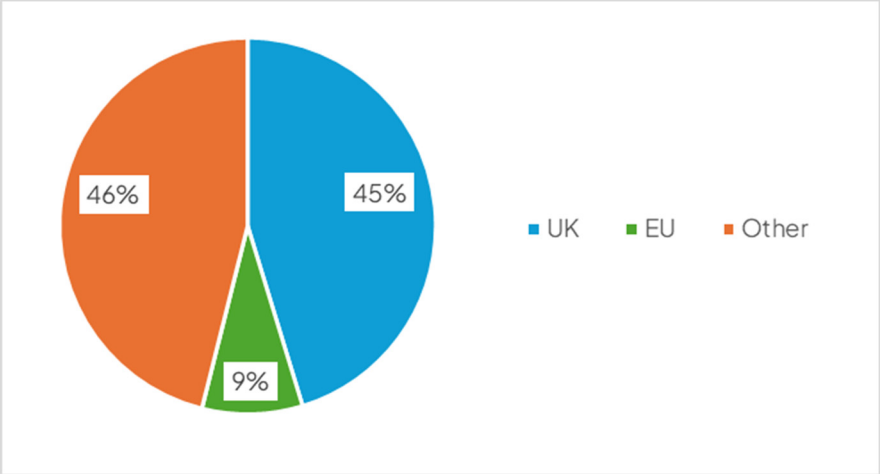
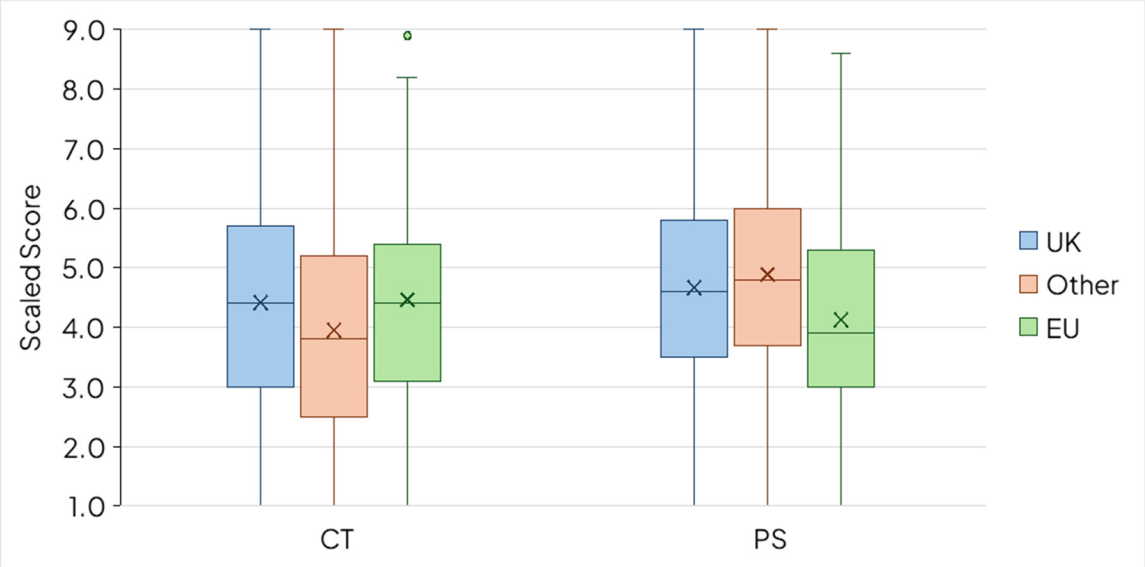


Table 7 summarises the scaled score distributions for the two modules and the trends are quite different as can be observed in the box and whisker plot in Figure 6. For Critical Thinking, the EU candidates were the strongest with a mean score of 4.46 compared to Other candidates who had a mean score of 3.95. This could be partly attributed to the higher linguistic load of this module. For the Problem Solving module, on the other hand, the Other students had the highest mean score at 4.89 compared to 4.13 for EU candidates which was the lowest scoring group. This is illustrated in the box and whisker plot in Figure 6.

Table 7 Scaled Score Summary Statistics by Area of Residence

Module	Area	*N*	*%N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Critical Thinking	UK	1,189	45%	4.41	1.89	1.0	9.0	3.0	4.4	5.7	6.7
	EU	227	9%	4.46	1.77	1	9	3.1	4.4	5.4	6.7
	Other	1,210	46%	3.95	1.95	1	9	2.5	3.8	5.2	6.7
Problem Solving	UK	1,189	45%	4.67	1.75	1.0	9.0	3.5	4.6	5.8	7.0
	EU	227	9%	4.13	1.78	1.0	8.9	3.0	3.9	5.3	6.3
	Other	1,210	46%	4.89	1.76	1.0	9.0	3.7	4.8	6.0	7.4

Figure 6 Box and Whisker Plot of Scaled Score by Area of Residence



4.2.4 Ethnicity (UK Candidates Only)

UAT UK candidates who reside in the UK are asked to describe their ethnicity. A detailed list of ethnicities is provided to the candidates, but for the purposes of this analysis, the groups are combined as listed in Table 8.

Table 8 Ethnic categories used for reporting

Ethnicity	Ethnic Group Used for Reporting
Arab	Other
Gypsy, Traveller or Irish Traveller	Other
Other	Other
Asian - Bangladeshi	Asian
Asian - Chinese	Asian
Asian - Indian	Asian
Asian - Other	Asian
Asian - Pakistani	Asian
Black - African	Black
Black - Caribbean	Black
Black - Other	Black
Other Mixed	Mixed
White and Asian	Mixed
White/ Black African	Mixed
White/ Black Caribbean	Mixed
White	White

Figure 7 shows the breakdown of candidates by ethnicity for the TARA modules. The largest ethnic group amongst UK candidates was Asian (50%), followed by White (23%). This is in contrast to the other UAT UK exams, where the largest UK group is White. This reflects the differing population for the TARA examination, which is primarily used for entry to University College London (UCL) as opposed to the University of Cambridge or Imperial University (plus others for TMUA). This may change again in the following cycle as more universities adopt this examination for admissions.

Figure 7 Percent of UK Candidates by Ethnicity

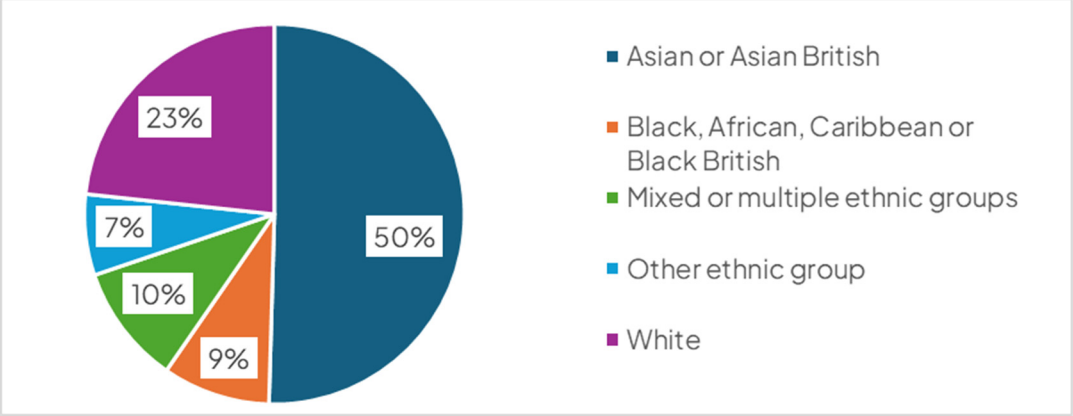


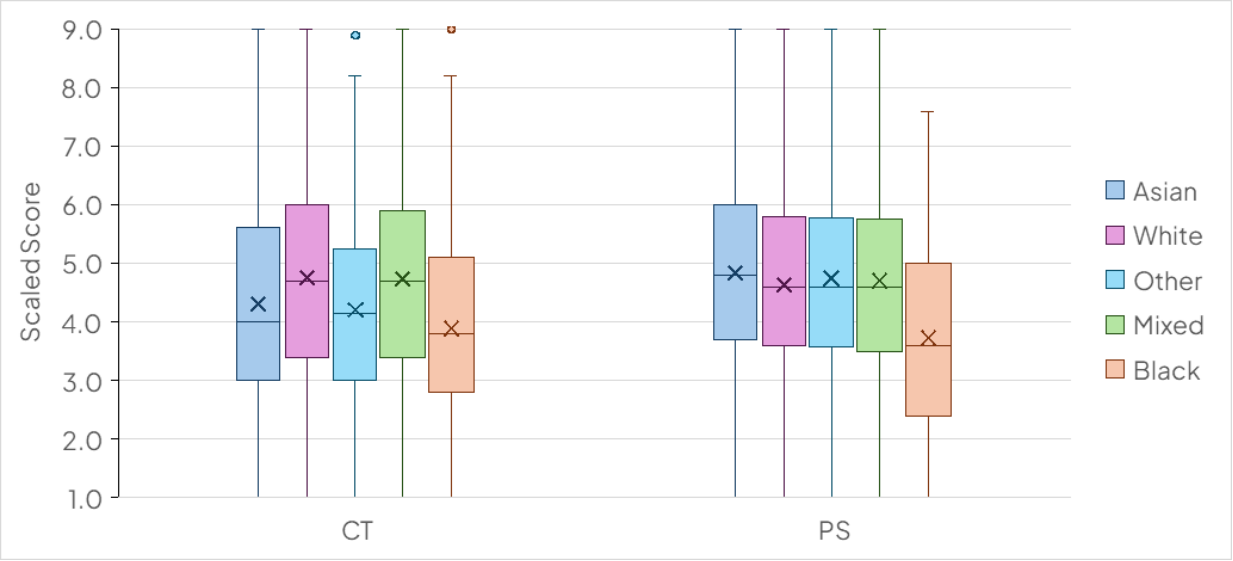
Table 9 summarises the scaled score summary statistics by ethnic group for UK students across the modules for groups with at least 50 candidates. The differences in scaled scores across the ethnic groups in the UK tend to reflect underlying trends within the UK.

Table 9 Scaled Score Summary Statistics by Ethnicity for UK Candidates

Module	Ethnicity	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Critical Thinking	Asian	600	50%	4.31	1.89	1.0	9.0	3.0	4.0	5.6	6.7
	Black	109	9%	3.89	1.77	1.0	9.0	2.8	3.8	5.1	6.4
	Mixed	121	10%	4.74	1.81	1.0	9.0	3.4	4.7	5.8	6.7
	Other	82	7%	4.20	1.80	1.0	9.0	3.0	4.2	5.2	6.4
	White	277	23%	4.76	1.93	1.0	9.0	3.4	4.7	6.0	7.6
Problem Solving	Asian	600	50%	4.84	1.79	1.0	9.0	3.7	4.8	6.0	7.4
	Black	109	9%	3.73	1.58	1.0	7.6	2.5	3.6	5.0	5.8
	Mixed	121	10%	4.70	1.66	1.0	9.0	3.5	4.6	5.7	6.7
	Other	82	7%	4.75	1.70	1.0	9.0	3.6	4.6	5.7	7.0
	White	277	23%	4.63	1.69	1.0	9.0	3.6	4.6	5.8	6.7

Figure 8 shows a box and whisker plot of scaled scores by ethnicity for each module. For Critical Thinking, UK White is the strongest performing group with a mean scaled score of 4.76, closely followed by UK Mixed at 4.74. The only group to have a mean scaled score below 4.0 was UK Black with a mean scaled score of 3.89. Interestingly, for Problem Solving, UK Asian was the strongest performing group, with a mean scaled score of 4.84, compared to a mean of 4.63 for UK White which was the third strongest group (after UK Mixed). As with Critical Thinking, the only group with a mean scaled score below 4.0 was UK Black with a mean scaled score of 3.73.

Figure 8 Box and Whisker Plot of Scaled Score by Ethnicity



4.2.5 First Language

Figure 9 illustrates the proportion of candidates who identified English as their first language for the TARA exam, showing an even split between those who listed English as their first language and those who did not.

Figure 9 Percentage of Candidates by First Language

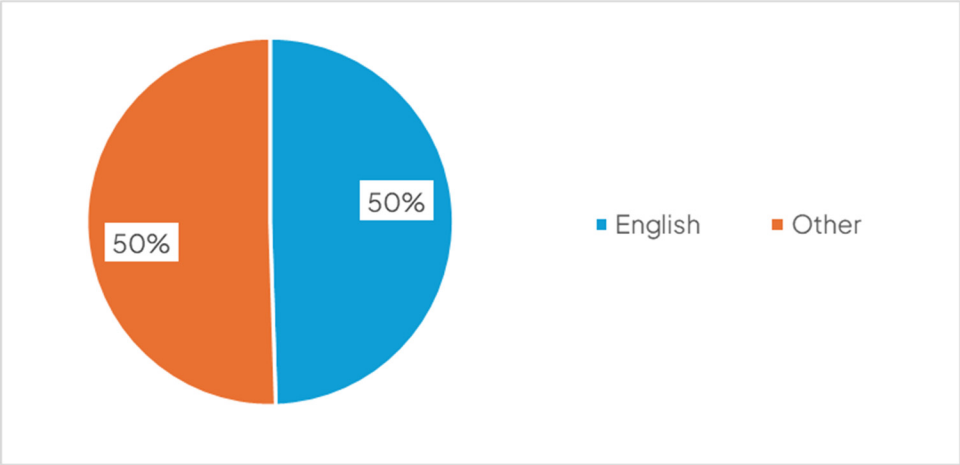
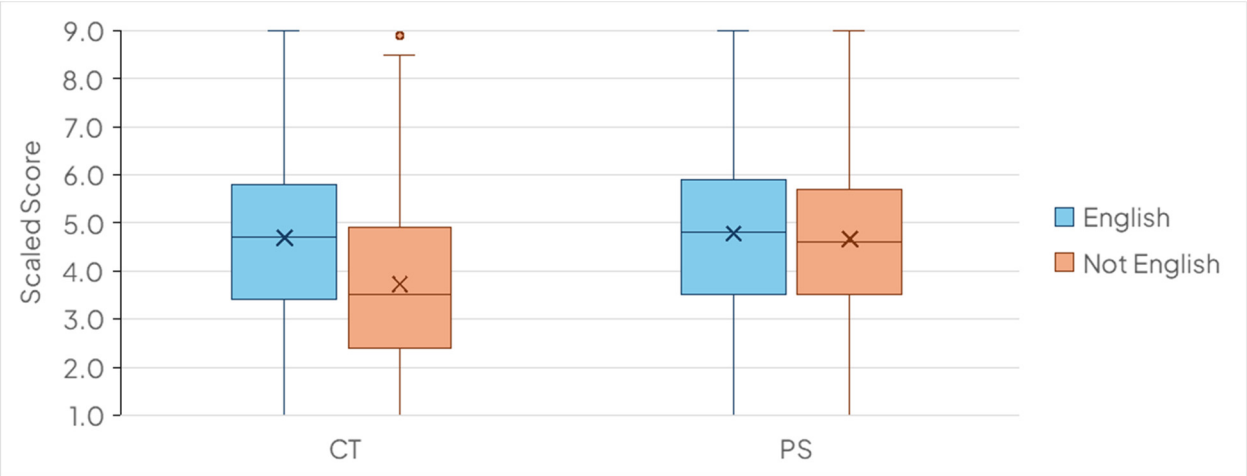


Table 10 Scaled Score Summary Statistics by First Language

Module	First Language	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Critical Thinking	English	1,300	50%	4.69	1.86	1.0	9.0	3.4	4.7	5.8	7.1
	Other	1,326	50%	3.73	1.87	1.0	9.0	2.4	3.5	4.9	6.4
Problem Solving	English	1,300	50%	4.78	1.78	1.0	9.0	3.5	4.8	5.9	7.0
	Other	1,326	50%	4.67	1.76	1.0	9.0	3.5	4.6	5.7	7.0

Figure 10 shows a box and whisker plot of scaled scores by first language for both modules. This illustrates that, for both modules, candidates who speak English as a first language outperformed those who did not. The difference in mean scaled scores between the two groups is much smaller for Problem Solving (0.11) than for Critical Thinking (0.95). This reflects the overall higher language load for Critical Thinking.

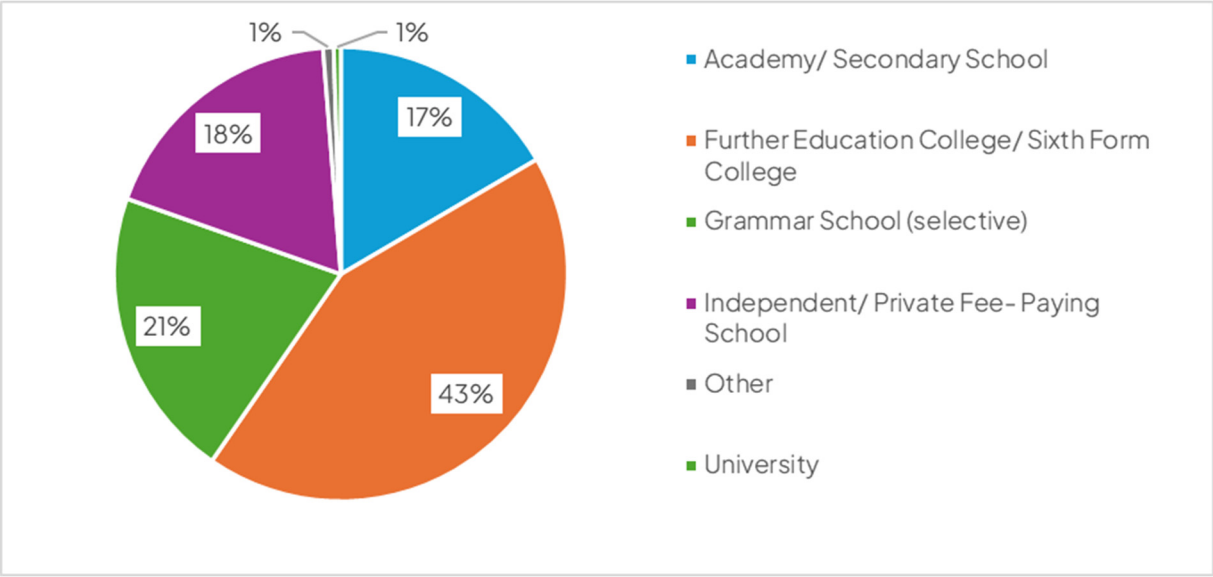
Figure 10 Box and Whisker Plot of Scaled Score by First Language



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 TARA was Further Education College or Sixth Form College (43%; Figure 11), with Academy/Secondary School (17%) being the least common school type in the UK, excluding Other and University. Grammar school candidates made up 21% of the UK candidates taking TARA, which is much higher than the national average (around 5% of state-funded secondary pupils attend grammar schools).

Figure 11 Percentage of Candidates by School Type

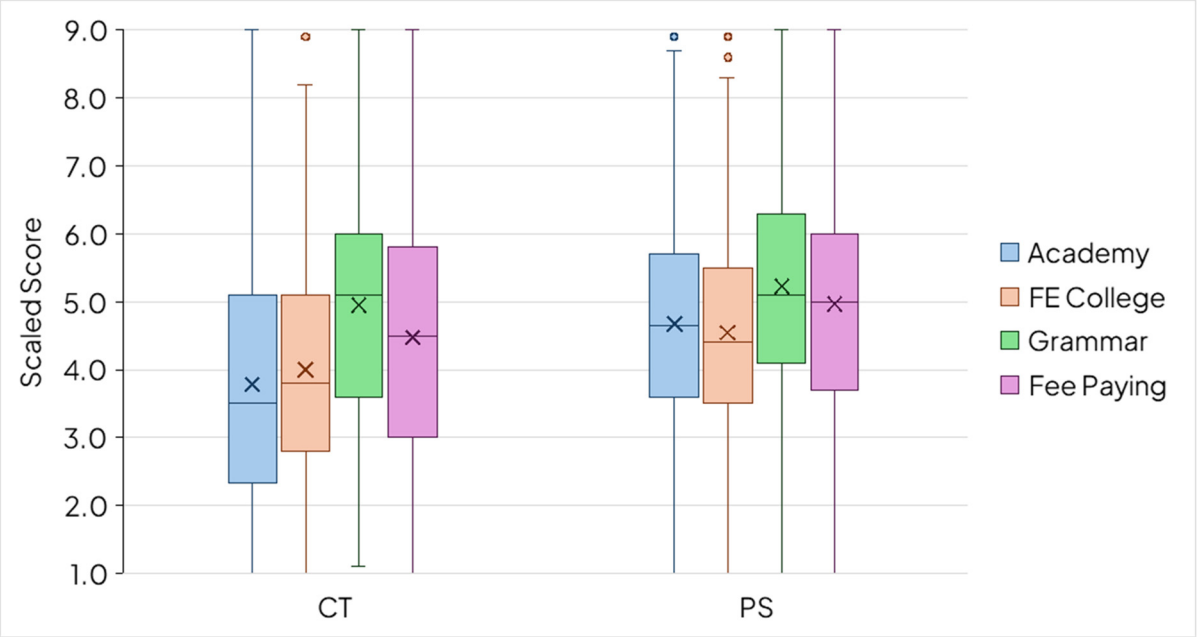


The scaled score summary statistics for groups with at least 100 candidates can be found in Table 11. Figure 12 shows these results in a box and whisker plot of scaled scores by school type for each module. The plot indicates that the four main school categories fall into two groups. Further Education (FE) Colleges and Academy/Secondary Schools had comparable scaled scores, with Academy/Secondary Schools having slightly higher scores than FE Colleges. Grammar and Private/fee-paying schools had overall higher mean scaled scores than non-selective state schools, with mean scaled scores from over 5.0 for grammar/private schools to below 4.5 for non-selective state schools. This is in line with national trends, where selective and private schools tend to outperform other types of schools.

Table 11 Scaled Score Summary Statistics by School Type

Module	School Type	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Critical Thinking	Secondary School	197	17%	4.20	1.88	1.0	9.0	2.8	4.0	5.4	6.7
	FE/ 6th Form College	512	43%	3.89	1.78	1.0	9.0	2.8	3.8	4.9	6.4
	Grammar School	247	21%	5.00	1.78	1.1	9.0	3.6	5.1	6.0	7.1
	Private Fee-Paying	218	18%	5.13	1.89	1.0	9.0	4.0	5.1	6.7	8.2
Problem Solving	Secondary School	197	17%	4.25	1.76	1.0	9.0	3.0	4.2	5.5	6.5
	FE/ 6th Form College	512	43%	4.37	1.67	1.0	9.0	3.2	4.4	5.4	6.5
	Grammar School	247	21%	5.25	1.75	1.0	9.0	4.1	5.3	6.3	7.6
	Private Fee-Paying	218	18%	5.08	1.71	1.0	9.0	3.9	5.0	6.0	7.6

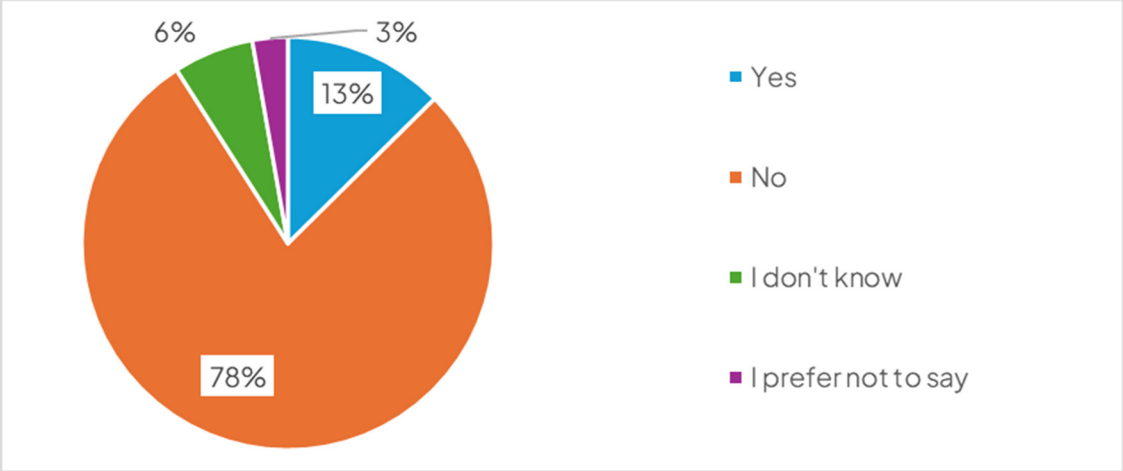
Figure 12 Box and Whisker Plot of Scaled Score by School Type



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. Typically, candidates who receive free school meals are from families with a lower income and can be classified as widening participation candidates. In total, 13% of UK candidates stated that they received free school meals (Figure 13) compared to almost 25% of pupils nationally who are eligible for free school meals. Note that the candidates were also given the options "I don't know" and "I prefer not to say" in addition to "Yes" and "No", which accounted for 9% of the UK candidates.

Figure 13 Percentage of Candidates by Free School Meals (UK Only)

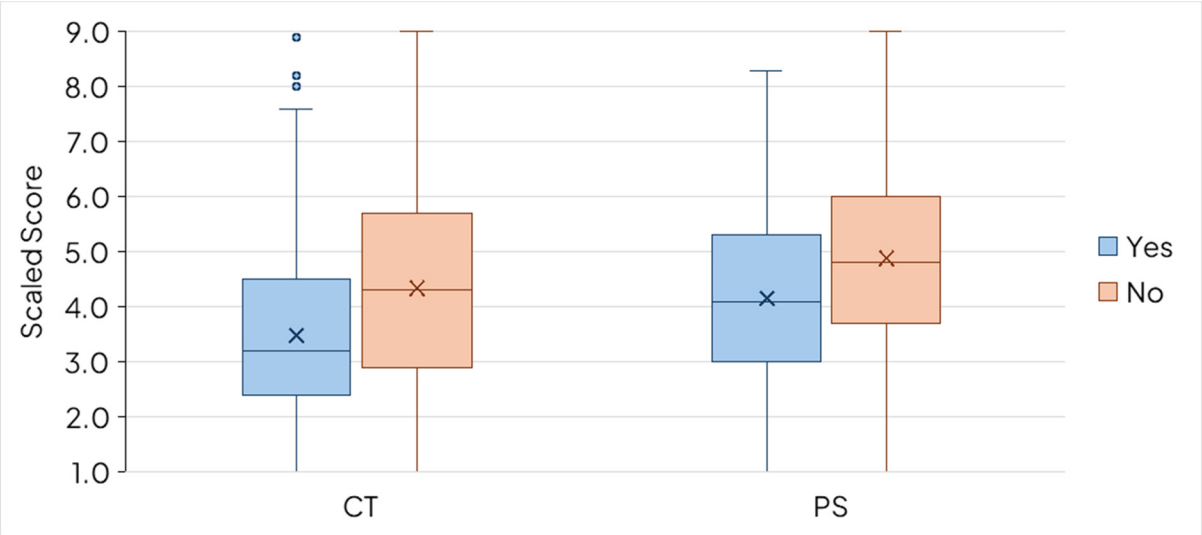


The scaled score summary statistics for the two main groups of candidates are summarised in Table 12 and are plotted in a box and whisker plot in Figure 14. This shows that the candidates who are receiving free school meals tend to have much lower scores than those who are or were not, which is in line with national trends. This is more marked for Critical Thinking than for Problem Solving.

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

Module	Free School Meals	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Critical Thinking	Yes	150	13%	3.76	1.59	1.0	8.9	2.8	3.6	4.9	5.8
	No	931	78%	4.56	1.92	1.0	9.0	3.1	4.5	5.8	7.1
Problem Solving	Yes	150	13%	4.03	1.74	1.0	8.9	2.9	3.9	5.3	6.2
	No	931	78%	4.78	1.74	1.0	9.0	3.6	4.6	5.9	7.0

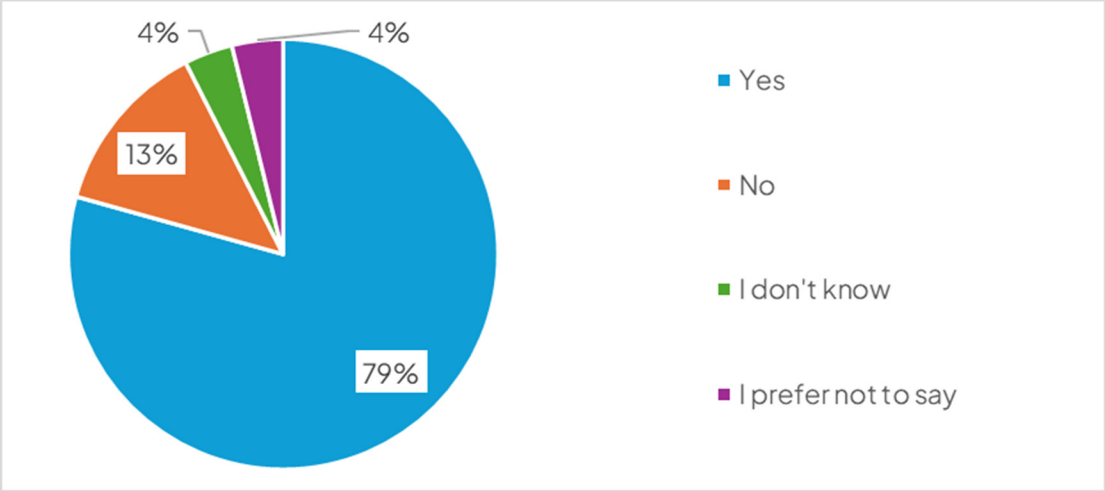
Figure 14 Box and Whisker Plot of Scaled Score by UK Free School Meals (Yes/No)



4.2.8 Parent Higher Education

All candidates were asked if their parent or guardian had attended tertiary education (that is, gained post-school qualifications). A total of 79% of candidates responded "Yes" to this question (Figure 15). Responding "No" to this question can be considered, along with free school meals, to be an indicator of widening participation candidates.

Figure 15 Percentage of Candidates by Parent Higher Education

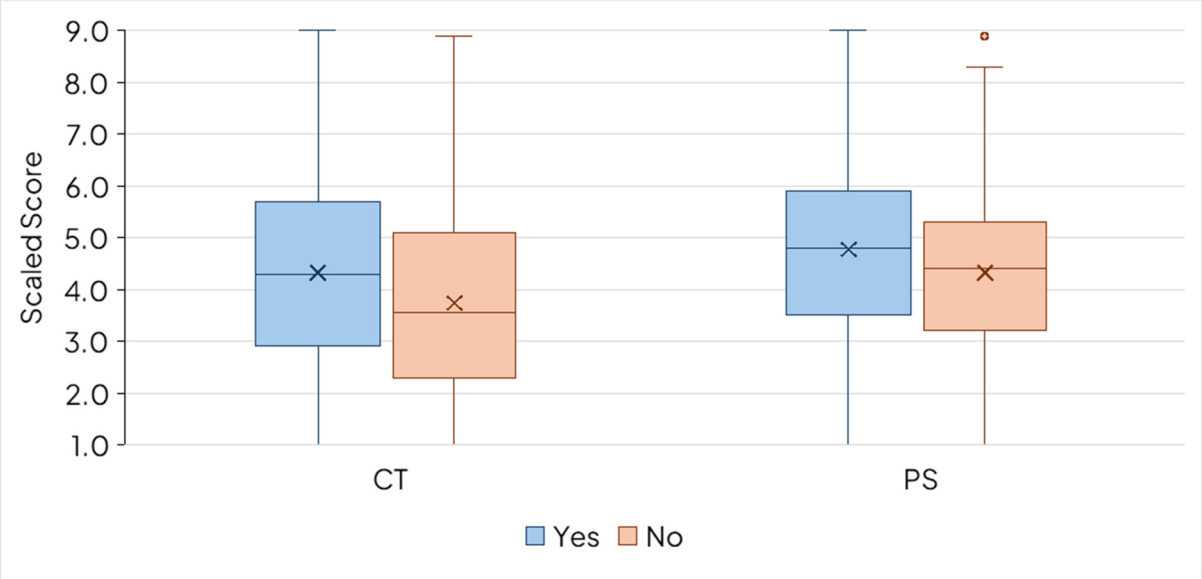


The scaled score summary statistics are shown in Table 13 and are illustrated in a box and whisker plot in Figure 16. These data show that across both modules, the candidates who had a parent/guardian who attended higher education significantly outperformed those who did not, although candidates who responded "I don't know" and "I prefer not to say" were also strong performers.

Table 13 Summary Statistics by Parent/Guardian Education

Module	Parent Higher Education	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Critical Thinking	Yes	2,083	79%	4.33	1.92	1.0	9.0	2.9	4.3	5.7	6.7
	No	346	13%	3.74	1.78	1.0	8.9	2.3	3.6	5.1	6.4
	I don't know	96	4%	3.68	1.89	1.0	9.0	2.4	3.5	4.8	6.7
	I prefer not to say	101	4%	3.71	2.07	1.0	9.0	2.1	3.4	5.4	6.4
Problem Solving	Yes	2,083	79%	4.78	1.78	1.0	9.0	3.5	4.8	5.9	7.1
	No	346	13%	4.32	1.70	1.0	9.0	3.2	4.4	5.3	6.5
	I don't know	96	4%	4.81	1.72	1.0	9.0	3.7	4.8	5.9	7.0
	I prefer not to say	101	4%	4.87	1.72	1.1	9.0	3.8	4.8	5.7	7.6

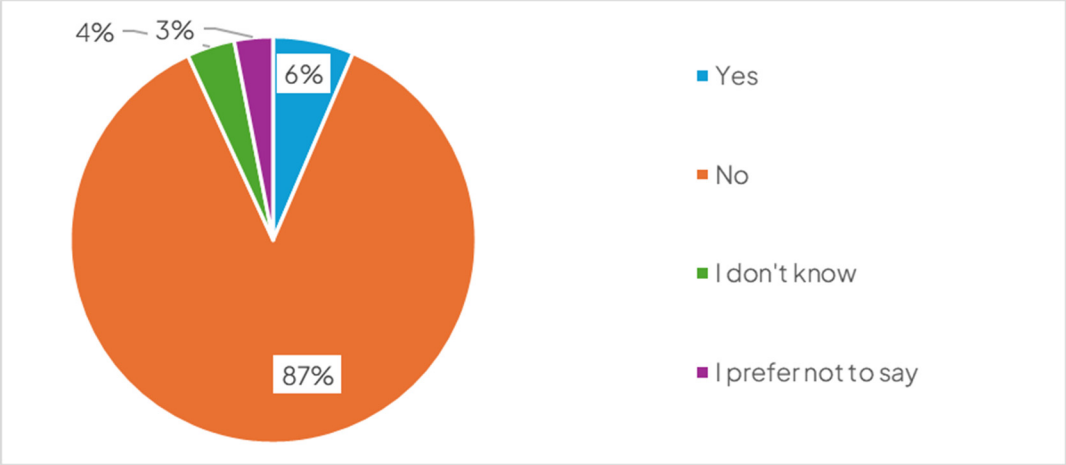
Figure 16 Box and Whisker Plot of Scaled Score by Parent/Guardian Higher Education



4.2.9 Learning Difficulty/Chronic Heath Condition

Candidates are asked whether they have a learning disability or health condition that has lasted (or is expected to last) for a year or longer. For TARA, 87% of candidates said that they did not have such a health condition or disability (Figure 17). It should be noted that candidates also had the option of "I prefer not to say" and "I don't know", which accounted for 7% of the candidates.

Figure 17 Percent of Candidates by Learning Difficulty/Chronic Health Condition

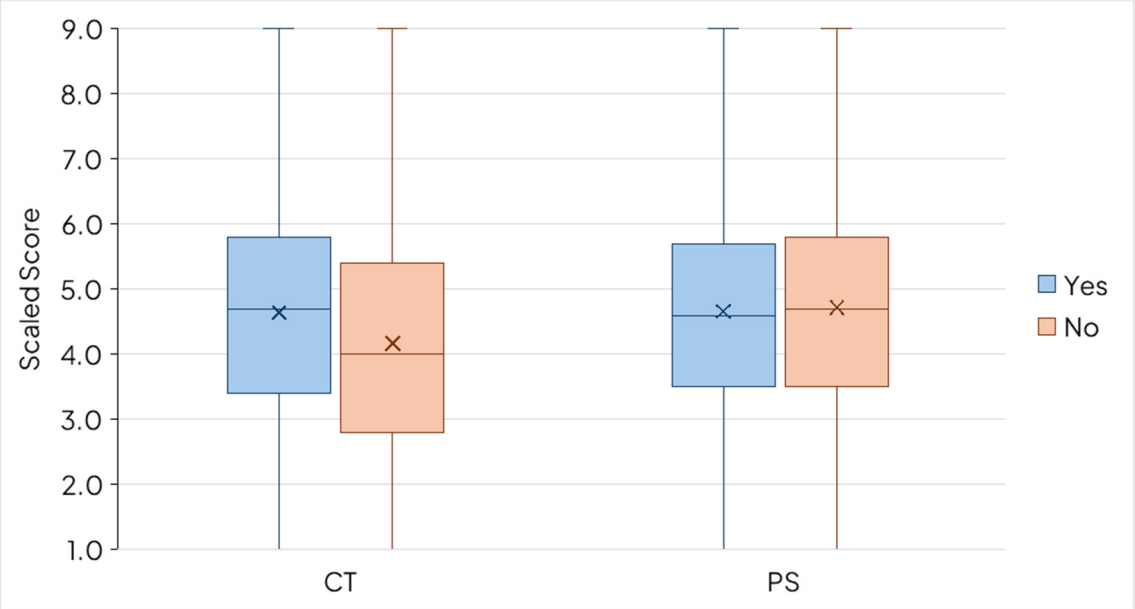


There is a difference in performance across the two groups, with those who responded "Yes" outperforming those who responded "No" in Critical Thinking. In Problem Solving, those who responded "No" outperformed those who responded "Yes" (Table 14; Figure 18) but candidates who chose to respond "I prefer not to say" also performed very strongly.

Table 14 Summary Statistics by Learning Difficulty/Chronic Health Condition

Module	Learning Difficulty	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Critical Thinking	Yes	169	6%	4.64	1.70	1.0	9.0	3.4	4.7	5.8	6.7
	No	2,276	87%	4.17	1.94	1.0	9.0	2.8	4.0	5.4	6.7
	I don't know	100	4%	4.35	1.74	1.1	8.9	3.1	4.3	5.6	6.7
	I prefer not to say	81	3%	4.12	2.11	1.0	9.0	2.4	4.1	5.7	7.0
Problem Solving	Yes	169	6%	4.66	1.80	1.0	9.0	3.5	4.6	5.7	6.7
	No	2,276	87%	4.72	1.77	1.0	9.0	3.5	4.7	5.8	7.0
	I don't know	100	4%	4.64	1.63	1.0	9.0	3.2	4.5	5.7	6.5
	I prefer not to say	81	3%	5.04	1.78	1.0	9.0	3.8	5.0	6.0	7.6

Figure 18 Box and Whisker Plot of Scaled Score by Learning Disability/Chronic Health Condition



4.3 Access Arrangements

Candidates can request access arrangements, such as extra time, if required for their test. In total, 102 candidates had access arrangements for the TARA, which is around 4% of the candidate population. Candidates can apply for more than one access arrangement. Table 15 summarises the number of access arrangements by the type required. The most common request was for extra time and/or pause-the-clock, with 101 candidates requesting this.

Table 15 Number of Access Arrangements by Type

Access Arrangement	*N*
Extra Time	66
Pause-the-clock	7
Extra time + pause-the-clock	28
Separate Room	4
Other	7

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 (α), 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 α 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 α 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 σ_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 TARA modules each have only 22 items, making them relatively short and therefore the reliabilities are expected to be lower than the other UAT UK examinations. A reliability of over 0.70 would still be considered satisfactory for the lengths of the TARA modules.

The raw score reliabilities for Critical Thinking were reasonable, ranging from 0.69 to 0.79. For Problem Solving, the reliabilities ranged from 0.59 to 0.75, with the reliability of 0.59 observed for one form in January looking to be an outlier, possibly partly due to the low numbers of candidates (263) from a restricted candidate population. The remaining forms range from 0.68 to 0.72, which is very homogeneous. The reliabilities for both modules may well improve as the volume of candidates increases on the exam, as currently the candidate volumes per form are still relatively low.

Table 16 Raw Score Reliability and SEM

Module	Event	Raw Score Reliability	
		Cronbach's Alpha	*SEM*
Critical Thinking	Oct 2025	0.69 to 0.79	1.81 to 1.98
	Jan 2026	0.69 to 0.72	2.01 to 2.12
Problem Solving	Oct 2025	0.69 to 0.75	1.92 to 2.00
	Jan 2026	0.59 to 0.72	1.97 to 2.00

As TARA candidates receive a scaled score for each module, which is scaled from the candidate theta, the reliability of the theta estimate is also important when assessing the scaled scores. The scaled score reliability and SEM are summarised in Table 17. The scaled score reliabilities range from 0.61 to 0.77 for Critical Thinking and 0.61 to 0.74 for Problem Solving.

Table 17 Scaled Score Reliability and SEM

Module	Event	Raw Score Reliability	
		Cronbach's Alpha	*SEM*
Critical Thinking	Oct 2025	0.61 to 0.77	0.94 to 1.24
	Jan 2026	0.69 to 0.72	0.87 to 0.98
Problem Solving	Oct 2025	0.69 to 0.74	0.95 to 0.98
	Jan 2026	0.61 to 0.73	0.94 to 0.97

5.2 Test Timing Analysis

The module time for each candidate is calculated by summing the item and review time for each item and candidate for the items in the module (that is, not the non-disclosure agreement or survey). The time limit for each module is 00:40:00. The module time summary statistics are shown in Table 18. Candidates with a time over 00:40:04 — approximately 4% per module — were excluded from analysis of module time as they are assumed to have an access arrangement (the number of these candidates is listed in Table 19). The mean time for each of the modules is just under 39 minutes for Critical Thinking and just under 40 minutes for Problem Solving. The median module time can sometimes be a better indication of average module time, as extreme times have a smaller impact on the median than the mean. The median is just below 40 minutes for Critical Thinking and at 40 minutes for Problem Solving. The distribution of candidates by module time is illustrated in Figure 19, which shows that over 85% of candidates used between 35 and 40 minutes.

Table 18. Test Time Summary Statistics

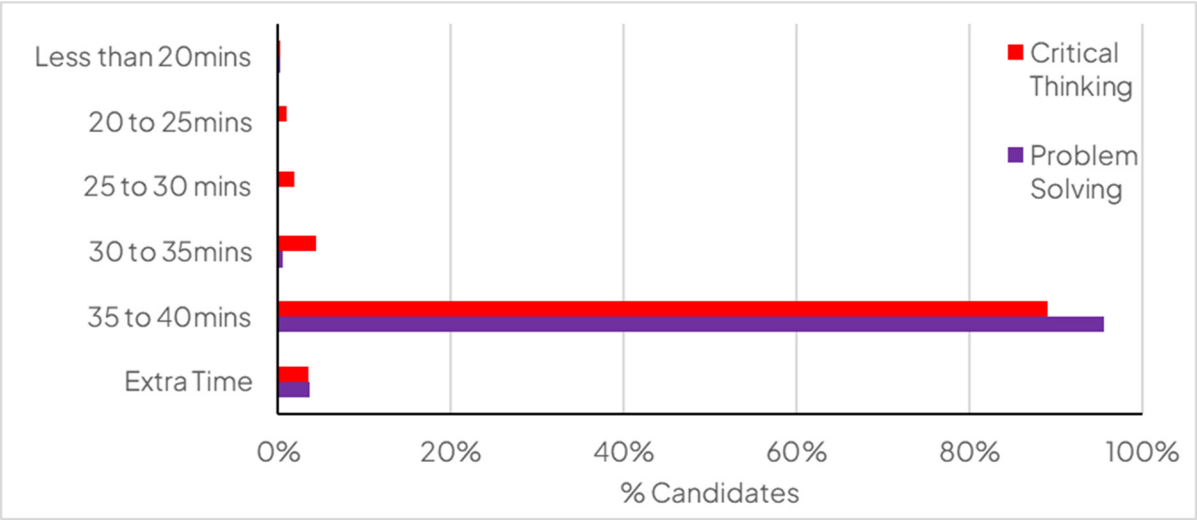
Module	*N*	Test Time				
		Mean	Median	*SD*	Min	Max
Critical Thinking	2,535	00:38:44	00:39:49	00:02:57	00:15:23	00:40:01
Problem Solving	2,532	00:39:43	00:40:00	00:01:25	00:06:04	00:40:01

Note: Candidates with extra time are excluded from this analysis.

Table 19 Candidates with Extra Time

Module	Test Time Greater Than 00:40:04	
	N	%
Critical Thinking	91	3%
Problem Solving	94	4%

Figure 19 Percentage of Candidates by Test Time



The test timing analysis implies that the majority of candidates are using the full test time available. This could suggest that the test is speeded. Speededness can be further assessed by looking at the number of unreached, or not presented, items. The TARA has non-linear navigation and therefore, within each module, candidates can choose the order in which they take the items; however, it is expected that most candidates will still choose to take the modules in a linear fashion. Therefore, any unreached items were likely not presented due to the candidate running out of time (or possibly ending the test early). The numbers of individual candidates with at least one not presented item are summarised in Table 20. This shows that for Critical Thinking only 1% of candidates had unreached items compared to 8% of Problem Solving candidates. It is typical for exams of this type to be slightly speeded. When admissions exams are well targeted in terms of difficulty and the ability to spread the top end of the ability range, this tends to lead to some candidates running out of time. This can be a trade-off when tests are targeted to a wide ability range.

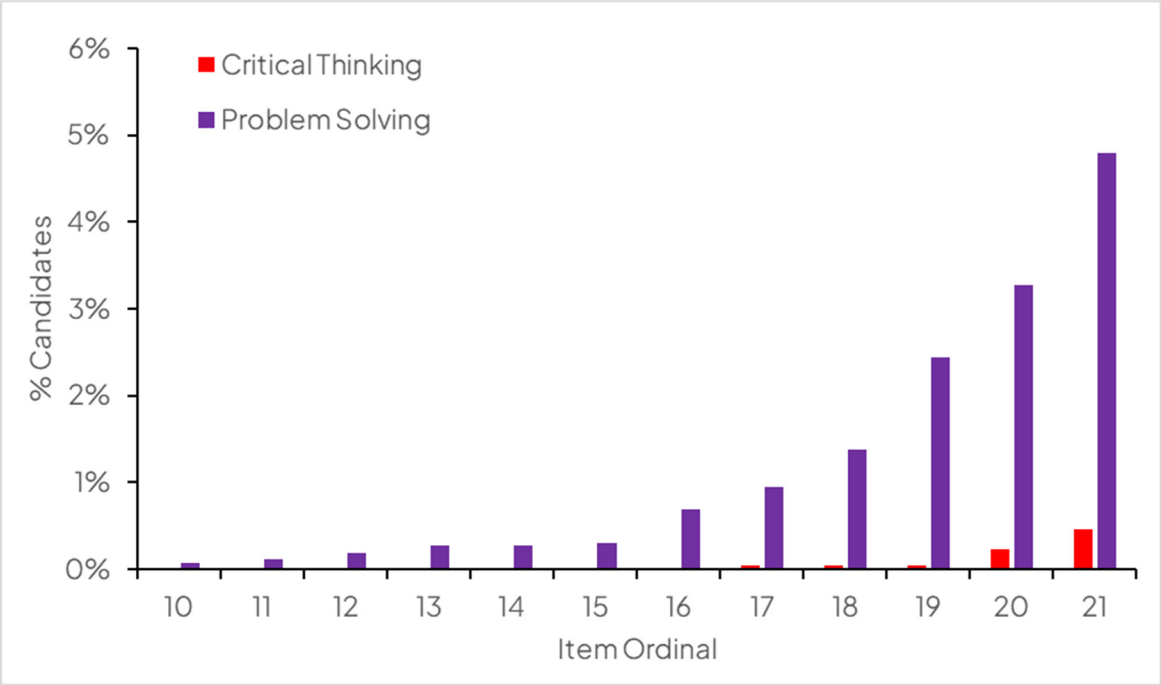
To reduce speededness further, item time could be considered when building forms in the future, as long as the supply of items is large enough to support this. Item time statistics should also be reviewed to identify any trends amongst items that are taking more time.

Table 20 Summary of Candidates with Unreached Items

Module	Total *N* Candidates	*N* Candidates with an Unreached Item	% Candidates with an Unreached Item	Mean *N* Items Unreached	Range
Critical Thinking	2626	33	1%	1.67	1 to 6
Problem Solving	2626	217	8%	2.76	1 to 12

The percentage of candidates with unreached items by item ordinal is plotted in Figure 23 (from Item 10). This plot includes all unreached items, so if a candidate did not reach items 20, 21, and 22, they are included for each of the three ordinals. There is, therefore, a cumulative element, as, in most cases, it can be assumed that candidates who do not reach item 25 will also not see items 26 and 27 if they are moving in a linear fashion. Figure 20 shows that Problem Solving is more speeded than Critical Thinking.

Figure 20 Percent of Unreached Items by Item Ordinal



In addition to candidates running out of time and having unreached items, it is also likely that, as candidates run out of time, they click quickly through the last few items and guess, as there is no negative marking. This information would not be captured in the unreached item analysis. Table 22 summarises the number of candidates with unreached or low-time items (below 5 seconds). The trends observed very much mirror those observed in Table 21, as the percentage of candidates with low time/unreached items is lower for Critical Thinking (4%) than Problem Solving (28%). This will continue to be monitored.

Table 21 Summary of Candidates with Unreached and Items of 5 Seconds or Less

Module	Total *N* Candidates	*N* Candidates with an Unreached/low Time Item	% Candidates with an Unreached/Low Time Item	Mean *N* Items Unreached/ Low Time	Range
Critical Thinking	2626	116	4%	2.23	1 to 9
Problem Solving	2626	742	28%	2.55	1 to 12

6. Item Performance

Each year, Pearson undertakes 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 TARA Item Analysis

For all TARA modules, item quality is assessed on three statistical criteria outlined below.

Point Biserial:
Item point biserial is the correlation between the item score and the test score, ranging from -1 to +1. A high value indicates that candidates who do well on the test do well on the item, and therefore the item discriminates well between strong and weak candidates. Items with a negative point biserial indicate that weak candidates are performing well on the item.

***p* Value**
The *p* value is 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. Note that for TARA, the forms are not all administered to the same candidate population, and therefore the *p* value is not always directly comparable, and caution should be taken when evaluating items.

IRT *b* Value
As outlined above, the item *p* value is not a reliable indicator of item difficulty when there are such different populations across the forms as there are for the UAT-UK tests. Therefore, in order to compare how similar the forms were in terms of difficulty, it is necessary to look at items on a common scale using IRT *b* values, which are on a single difficulty scale. Item *b* values typically range from -3.0 to +3.0, with harder items having higher or more positive values. As TARA is an admissions test, the aim of the test is to rank very able candidates to allow universities to select the most suitable applicants. Therefore, a wider range of difficulties is required for this type of test than for a traditional competency-based exam. There are also a range of courses using the test that will be selecting at different points along the scale, so the test needs to include both more accessible items and very challenging questions.

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 item writer training as examples of items that do not perform as expected.

6.1.1 Critical Thinking Item Performance

Of the items used in the Critical Thinking forms, 100% of items met the criteria. This is well above the 80% typical target for new items. All items had a positive point biserial, ranging from 0.10 to 0.57, with a mean of 0.38, which is excellent.

Table 22 Critical Thinking Item Status Outcome

Status	Comment	% of Items
Fail	*b* Value greater than +3 (difficult item)	0%
	b Value less than -3 (easy item)	0%
	Very low or negative point biserial	0%
Pass	*p* Value greater than 0.90	2%
	p Value less than 0.20	0%
	None	95%
Total		100%

6.1.2 Problem Solving Item Performance

Of the items used in the Problem Solving forms, 100% of items met the criteria. This is well above the 80% typical target for new items. All items had a positive point biserial ranging from 0.07 to 0.58, with a mean of 0.36, which is excellent.

Table 23 Problem Solving Item Status Outcome

Status	Comment	% of Items
Fail	*b* Value greater than +3 (difficult item)	0%
	b Value less than -3 (easy item)	0%
	Very low or negative point biserial	0%
Pass	*p* Value greater than 0.90	4%
	p Value less than 0.20	4%
	None	92%
Total		100%

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 the TARA modules, 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

Critical Thinking

No significant DIF was identified by Gender, First Language, UK School Type, or UK Ethnicity for Critical Thinking.

Problem Solving

The DIF results are reported in Table 24 for Problem Solving 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 flagged as showing DIF by gender, with this item favouring candidates who identified as a "Man" over those who identified as a "Woman". This item will be reviewed to identify likely sources of bias, and this information used to inform future item writing.

Table 24 Problem Solving Items Flagged with C Category DIF

Category	Item	*b* Value	Group Preferred	MH DIF Value	*p* Value (significance < 0.001)
Gender	820707	-0.7475	Man	1.73	0.0004

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.