University Admissions Tests UK

UAT-UK Ltd

Annual Reports 2024-2025

ESAT 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 Engineering and Science Admissions Test (ESAT) was administered in two windows: 15th and 16th October 2025, and 7th and 8th January 2025, for candidates applying to start university in 2025. The test is modular, with all candidates taking Mathematics 1 and then a selection of the remaining modules (Mathematics 2, Physics, Biology and Chemistry). This report covers the 11,919 ESAT candidates (9,141 in October 2024 and 2,778 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 science and mathematics-based exams. On average, male candidates tended to outperform female candidates across all modules. This is only counter to national trends in Biology where female candidates outperform male candidates at A-level, which is not observed here. Candidates with higher socio-economic status generally performed better than those with widening participation indicators. 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. Candidates who identified as Chinese significantly outperformed UK nationals across all modules, although it should be noted that these candidates represent a highly able sub-set of their cohort who are in a position to apply to competitive overseas universities. These differences were largest in the Mathematics modules.

Multiple forms (versions) of the ESAT modules were used, and these were administered at different times to different regions. The modules are relatively short, with 27 items, and therefore the reliability would be expected to be lower than for a longer test. Given this, the reliabilities for the ESAT tests were satisfactory, with Mathematics 1 showing the highest values. The forms were reasonably well balanced in terms of difficulty despite no item statistics being available when the items were selected from the forms.

The candidates taking ESAT include very able students, so the test is designed to challenge even the best candidates and includes difficult questions. 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 module. Biology was the least speeded module and Mathematics 2 was the most speeded, this might be expected as this module was overall too difficult for many candidates.

The items used in the ESAT 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 in each module 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.

2. Introduction

The Engineering and Science Admissions Test (ESAT) is designed to support universities in identifying strong applicants to degree courses related to engineering and science. It is used by a number of universities to distinguish between a large number of strong candidates with similar academic profiles.

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 ESAT is available in two sittings per admissions cycle. This report covers those candidates who took the test in October 2024 (15th and 16th) and January 2025 (7th and 8th). The test consists of five modules: Mathematics 1, Mathematics 2, Chemistry, Biology and Physics. Candidates are all required to take Mathematics 1 plus one to three other modules depending on the courses they are applying for. For each module the candidate receives a scaled score from 1.0 to 9.0 with no overall score. Section 3 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) 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 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 2024/2025 ESAT contains five 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 ESAT Test Design

Test	Module	Who takes this module?	Questions	Duration
ESAT	Mathematics 1 (compulsory)	All candidates take this module	Each module has 27 multiple-choice questions	Each module is 40 minutes
	Biology	Most candidates take two or three of these modules depending on the university courses they are applying for		
	Chemistry			
	Physics			
	Mathematics 2			

Candidates are given 40 minutes per module to answer a total of 27 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 reported to one decimal place for each module. 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 in Chemistry, for example, in either January or October has a higher ability than a candidate who scored 4.2 in Chemistry in either event. The scaled scores are not comparable across modules and there is no aggregate or total score. Further details of the scoring process are provided in the subsequent sections.

3.2 Measurement Model

3.2.1 Item Analysis

The five ESAT modules are all treated as separate tests for the analysis. For the 2024/2025 cycle, none of these items 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 itemresponse 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 ESAT 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 compare candidates, which is the prime objective of an admissions test. For each ESAT module, the candidates are issued a scaled score ranging from 1.0 to 9.0 reported to one decimal place.

The ESAT 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. 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 ESAT Modules

Event	Module	Percentile	Ability	Scaled Score
Oct 2024	Maths 1	50	0.3669	4.5
		90	2.1081	7.0
	Biology	50	0.4241	4.5
		90	1.6223	7.0
	Chemistry	50	0.2069	4.5
		90	1.6202	7.0
	Physics	50	0.0070	4.5
		90	1.3933	7.0
	Maths 2	50	-0.3466	4.5
		90	0.9266	7.0

Table 3 Scaling Constants

Module	Constant	Multiplier
Maths 1	3.9732	1.4358
Biology	3.6151	2.0865
Chemistry	4.1340	1.7689
Physics	4.4874	1.8034
Maths 2	5.1806	1.9636

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. ESAT is a modular test, with all candidates taking Maths 1 and then taking a number of other modules. The distribution of module combinations is shown in Table 4 with Maths 1, Maths 2 and Physics being the most common combination, taken by 72% of the candidates across the two events.

Table 4 ESAT Module Combinations

Combination	*N*	%
Maths 1, Maths 2, Physics	8,564	72%
Maths 1, Biology, Chemistry	1,201	10%
Maths 1, Maths 2, Chemistry	1,043	9%
Maths 1, Chemistry, Physics	566	5%
Maths 1, Maths 2	276	2%
Maths 1, Biology, Maths 2	153	1%
Other	116	1%

There were 11,919 ESAT candidates in total (all candidates take Maths 1), with 9,141 (77%) sitting in October 2024 and 2,778 (23%) in the January 2025 event (Table 5). 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 by module in Table 5. The individual modules are scaled separately and therefore the scaled scores are not directly comparable and cannot be aggregated.

The scaled score distribution is illustrated in Figure 1 to Figure 5 for each of the modules. Candidate scaled scores are normally distributed across the scaled score range, enabling universities to effectively differentiate between candidates.

Table 5 Scaled Score Summary Statistics

Module	Event	*N*	Mean	*SD*	Min	Max	Percentile			
							25	50	75	90
Maths 1	All	11,919	4.46	1.71	1.0	9.0	3.3	4.2	5.4	6.7
	Oct	9,141	4.59	1.76	1.0	9.0	3.4	4.5	5.7	7.0
	Jan	2,778	4.03	1.43	1.0	9.0	3.1	4.0	5.0	5.9
Biology	All	1,429	4.71	1.74	1.0	9.0	3.6	4.5	5.9	7.0
	Oct	1,428	4.72	1.74	1.0	9.0	3.6	4.5	5.9	7.0
	Jan	1	NA	NA	NA	NA	NA	NA	NA	NA
Chemistry	All	2,821	4.62	1.73	1.0	9.0	3.4	4.5	5.7	6.8
	Oct	2,570	4.70	1.73	1.0	9.0	3.4	4.5	5.8	7.0
	Jan	251	3.80	1.50	1.0	8.2	2.6	3.7	4.7	5.8
Physics	All	9,237	4.50	1.67	1.0	9.0	3.4	4.4	5.6	6.8
	Oct	6,853	4.65	1.70	1.0	9.0	3.4	4.5	5.6	7.0
	Jan	2,384	4.08	1.51	1.0	9.0	3.0	3.9	4.9	6.0
Maths 2	All	10,040	4.52	1.67	1.0	9.0	3.4	4.5	5.5	6.8
	Oct	7,303	4.66	1.75	1.0	9.0	3.6	4.5	5.6	7.0
	Jan	2,737	4.15	1.40	1.0	9.0	3.2	4.2	5.0	6.0

Figure 1 Mathematics 1 Binned Scaled Score Distribution

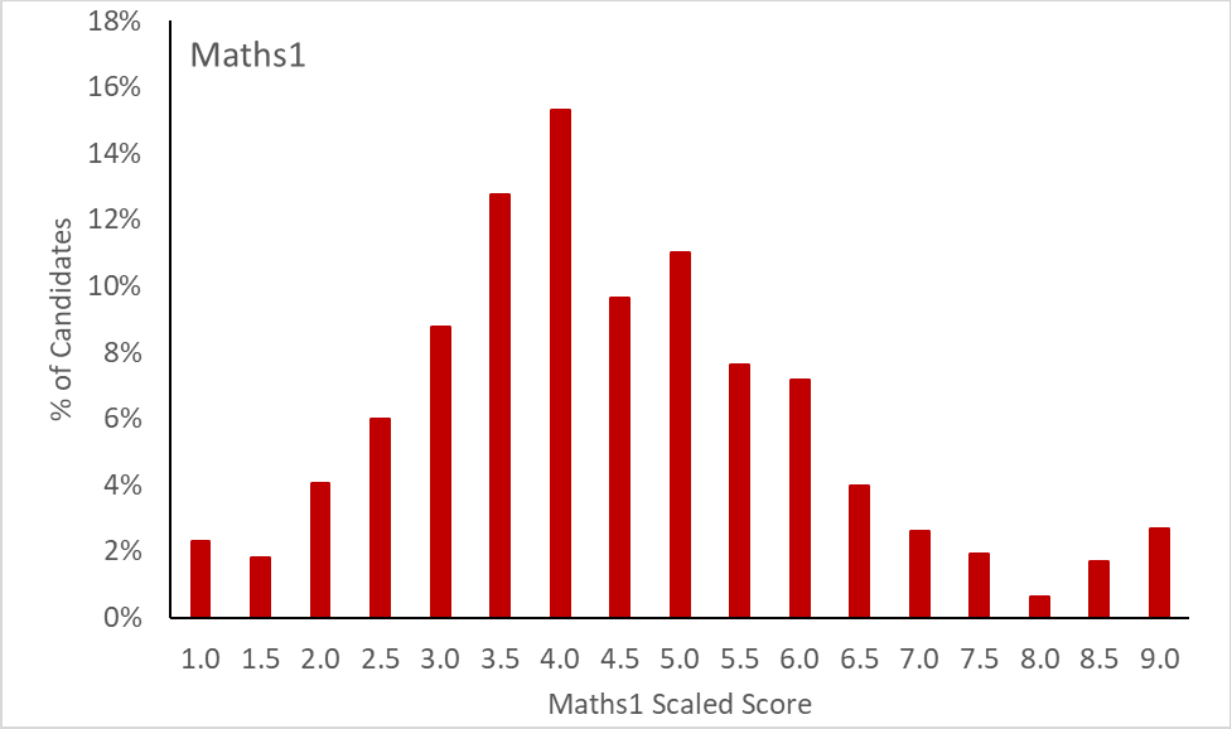


Figure 2 Biology Binned Scaled Score Distribution

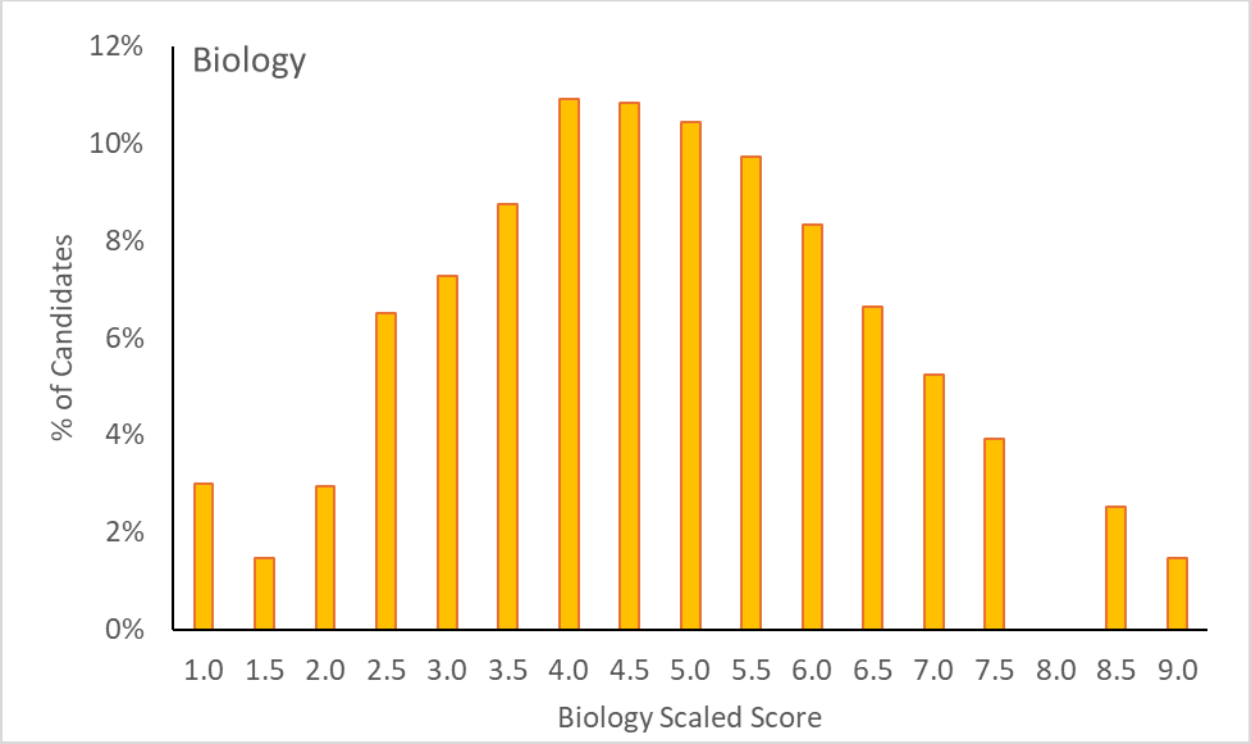


Figure 3 Chemistry Binned Scaled Score Distribution

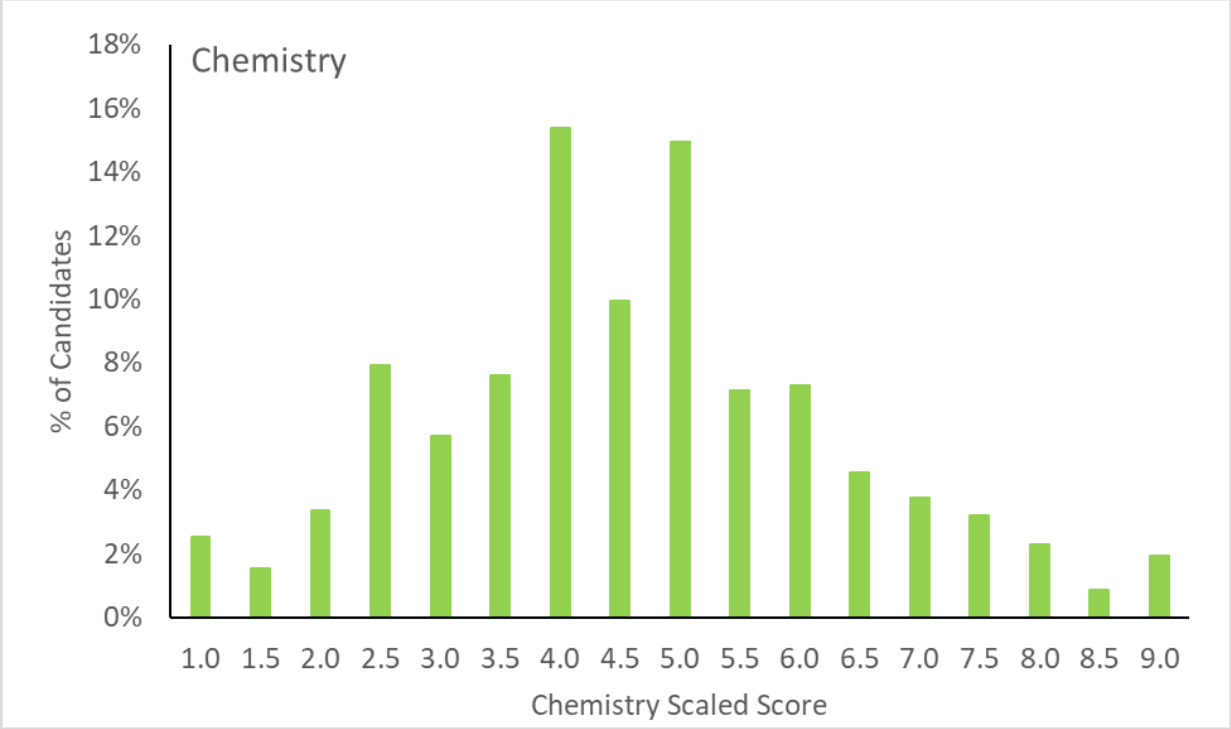


Figure 4 Physics Binned Scaled Score Distribution

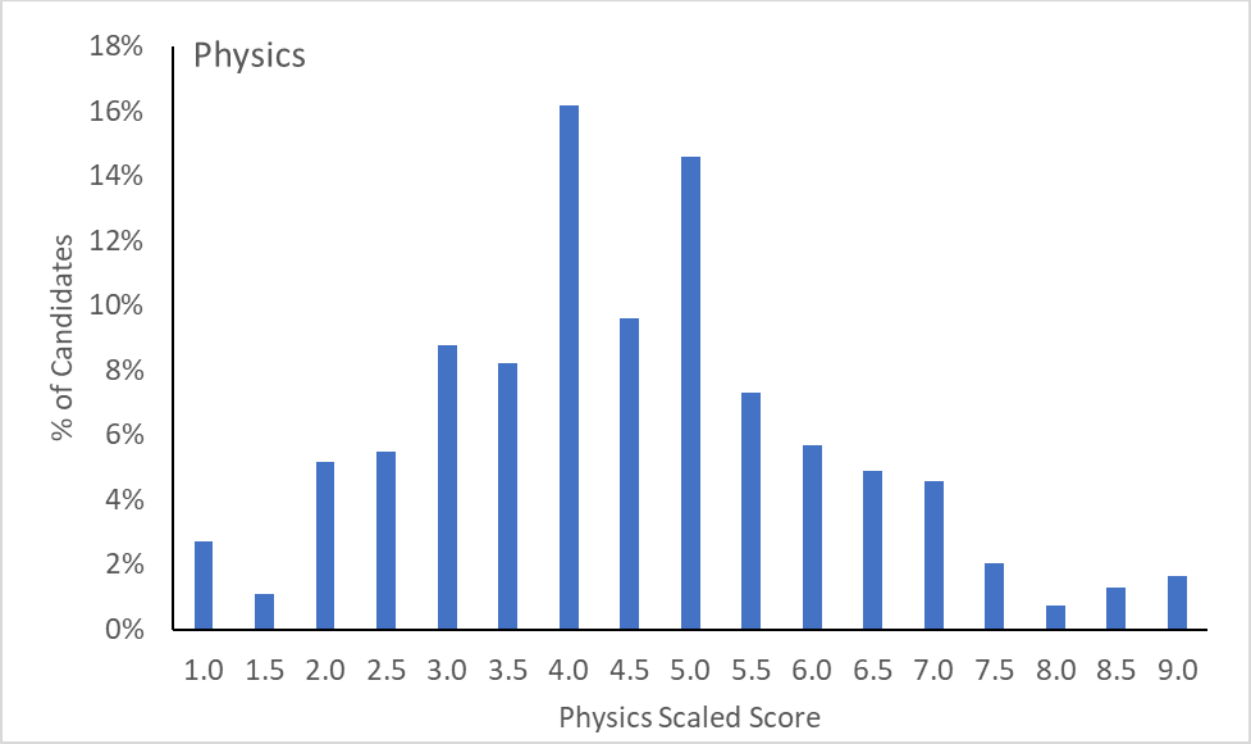
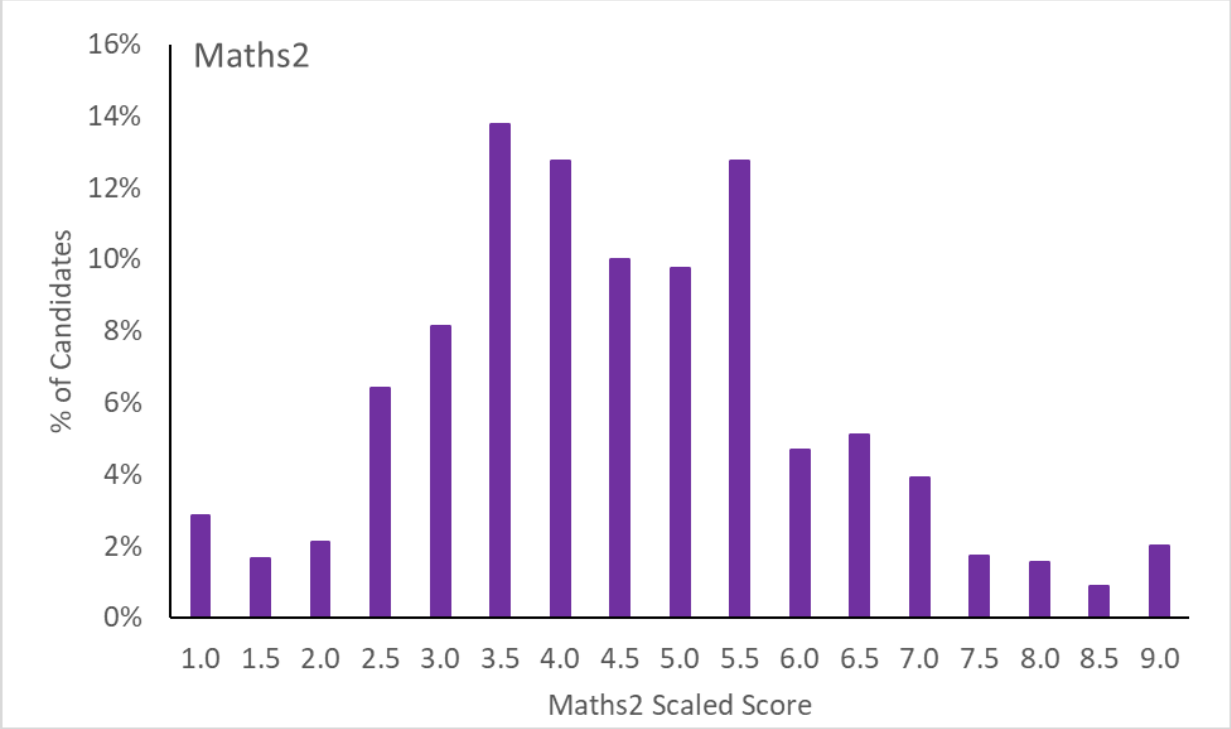


Figure 5 Mathematics 2 Binned Scaled Score Distribution



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 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 6.

Table 6 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 6 presents the breakdown of test-takers by gender across the five modules. This shows considerable variability across the different subjects. In the full cohort (i.e. Maths 1), 68% identified as "Man" and 31% as "Woman". Biology has the highest proportion of female candidates at 59% compared to only 25% for Physics.

Male candidates outperformed female candidates across all five modules and gained higher mean and median scaled scores (Table 7). The difference in mean scaled score ranges from 0.62 for Chemistry to 0.25 for Maths 2. The other mean differences were 0.41 for Maths 1, 0.47 for Biology and 0.50 for Physics. These differences are mostly in line with national A-level data except for Biology where female candidates typically outperform male candidates at A level.

Figure 6 Percent of Candidates by Gender

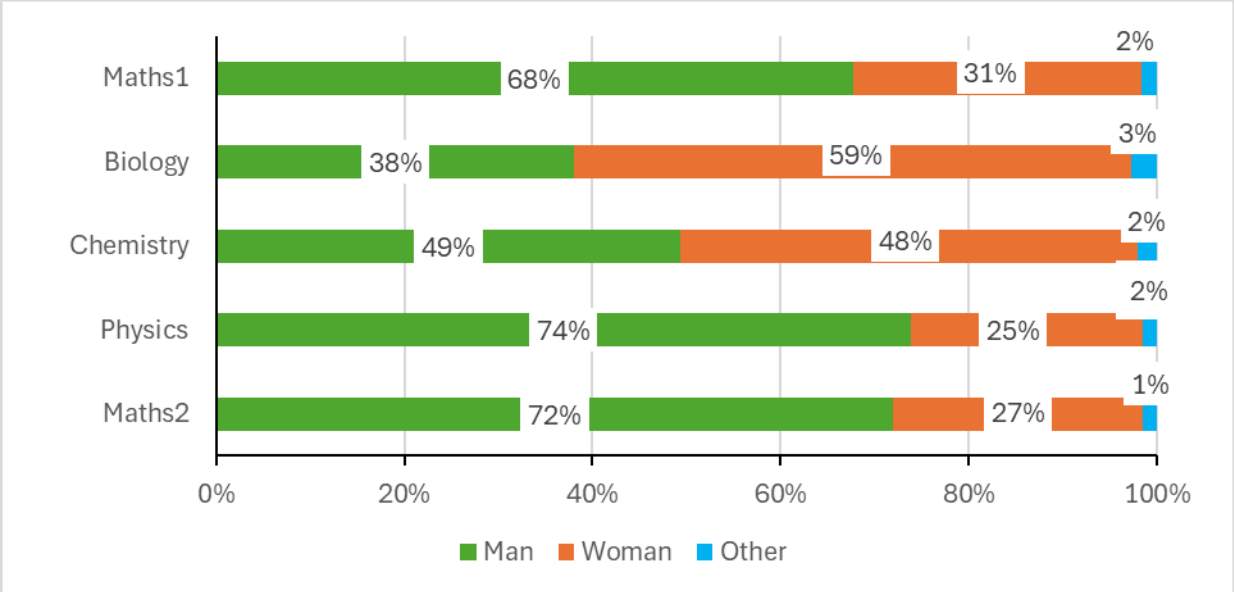
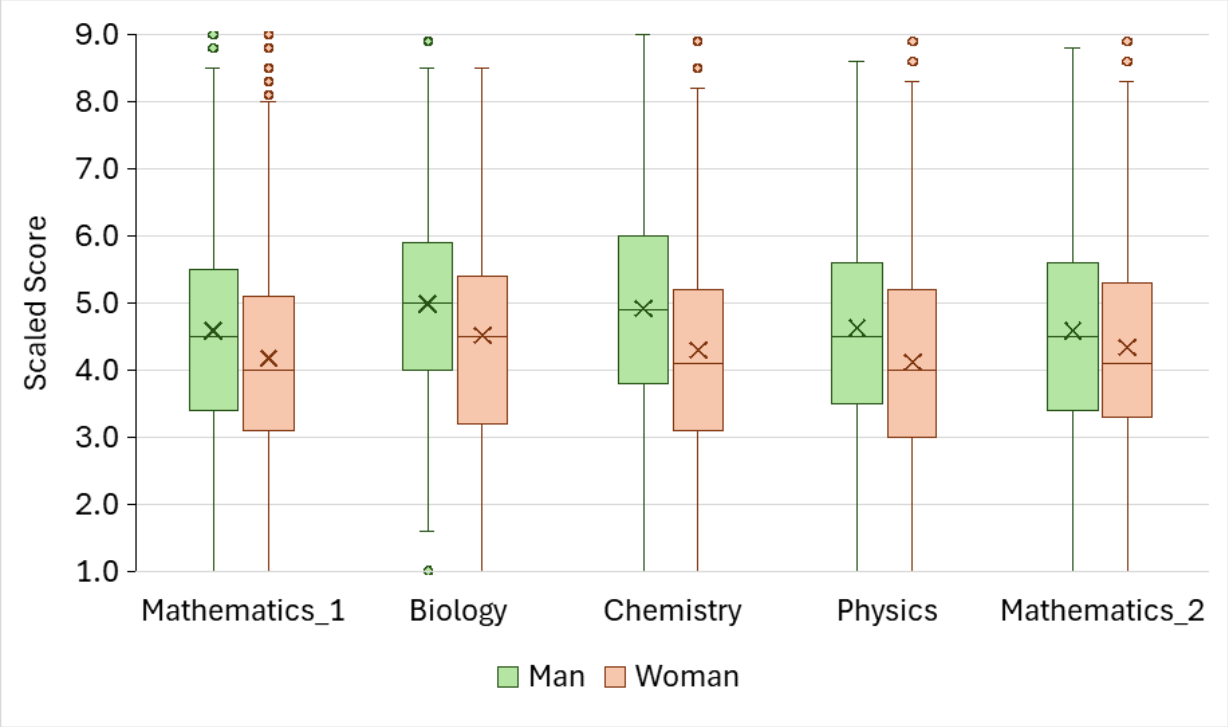


Table 7 Scaled Score Summary Statistics by Gender

Module	Gender	*N*	Scaled Score				Percentile			
			Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Man	8,074	4.59	1.69	1.0	9.0	3.4	4.5	5.5	7.0
	Woman	*3,646*	*4.18*	*1.71*	*1.0*	*9.0*	*3.1*	*4.0*	*5.1*	*6.5*
Biology	Man	544	4.99	1.68	1.0	9.0	4.0	5.0	5.9	7.0
	Woman	*845*	*4.52*	*1.75*	*1.0*	*9.0*	*3.2*	*4.5*	*5.4*	*7.0*
Chemistry	Man	1,393	4.92	1.73	1.0	9.0	3.8	4.9	6.0	7.3
	Woman	*1,368*	*4.30*	*1.68*	*1.0*	*9.0*	*3.1*	*4.1*	*5.2*	*6.5*
Physics	Man	6,817	4.62	1.65	1.0	9.0	3.5	4.5	5.6	6.8
	Woman	*2,275*	*4.12*	*1.67*	*1.0*	*9.0*	*3.0*	*4.0*	*5.2*	*6.3*
Maths 2	Man	7,224	4.59	1.67	1.0	9.0	3.4	4.5	5.6	6.8
	Woman	*2,668*	*4.34*	*1.69*	*1.0*	*9.0*	*3.3*	*4.1*	*5.3*	*6.4*

Figure 7 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 all modules.

Figure 7 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 are categorised as UK, EU or Other (Rest of World). Most ESAT candidates — 49% (Maths 2) to 55% (Chemistry) — were based in the UK, with only a small percentage in the EU and most of the remaining candidates outside of the UK and EU (Figure 8). The scaled score summaries can be found in Table 8 and are plotted in a box and whisker plot in Figure 9. This shows that UK and EU candidates were broadly similar in ability, but candidates from outside the European region were markedly stronger across most modules. This gap is slightly less marked for Biology (UK mean 4.64; Other mean 4.86) compared to Maths 1 (UK mean 3.93; Other mean 5.22) and Maths 2 (UK mean 4.07; Other mean 5.11).

Figure 8 Percent of Candidates by Area of Residence

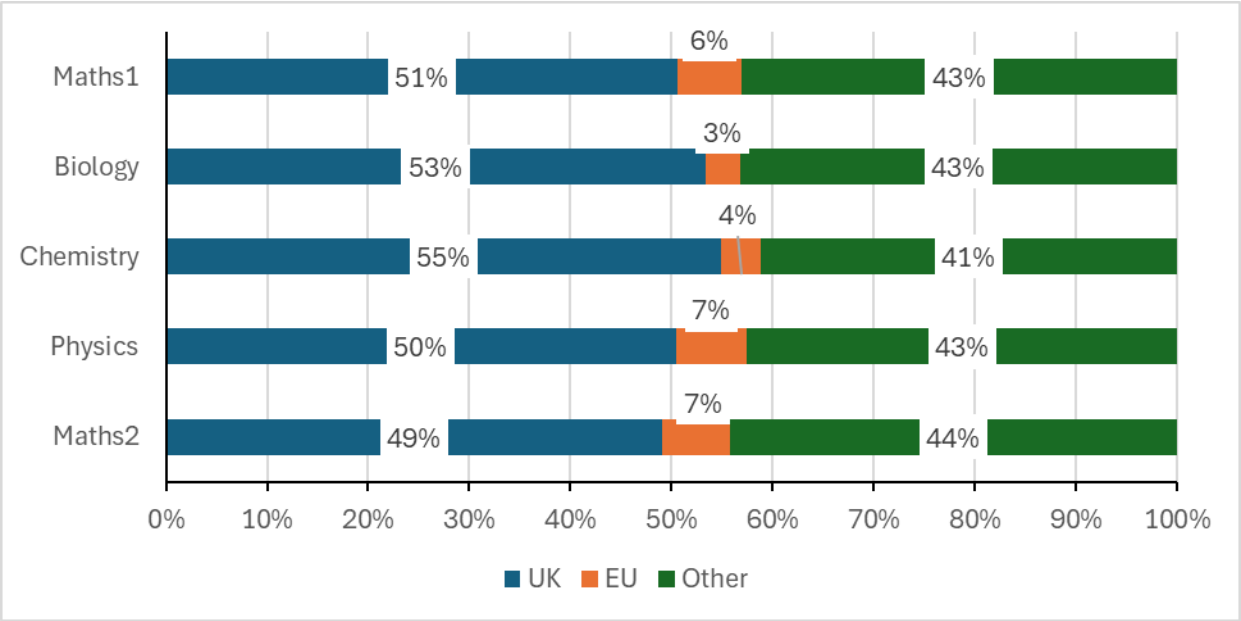
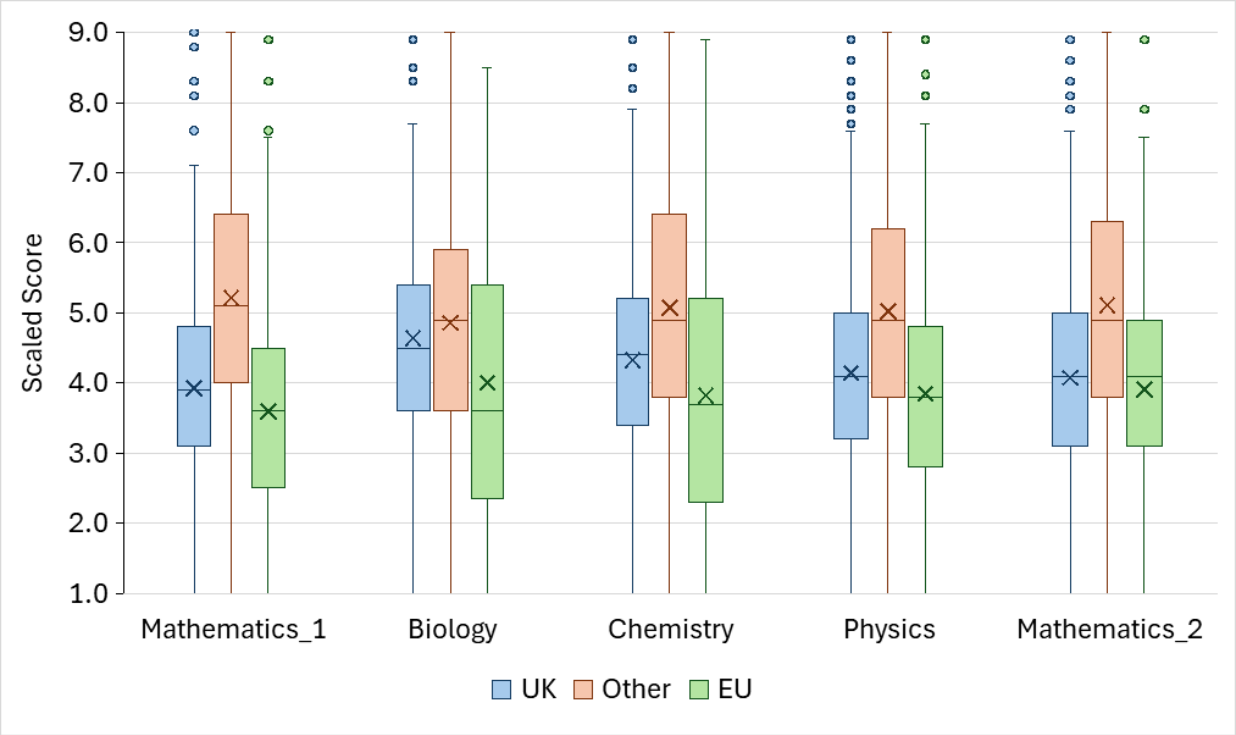


Table 8 Scaled Score Summary Statistics by Area of Residence

Module	Area	*N*	Scaled Score				Percentile			
			Mean	*SD*	Min	Max	25	50	75	90
Maths 1	UK	6,031	3.93	1.35	1.0	9.0	3.1	3.9	4.8	5.6
	EU	751	3.59	1.43	1.0	8.9	2.5	3.6	4.5	5.5
	Other	5,137	5.22	1.82	1.0	9.0	4.0	5.1	6.4	8.0
Biology	UK	762	4.64	1.67	1.0	9.0	3.6	4.5	5.4	7.0
	EU	49	4.00	1.94	1.0	8.5	2.4	3.6	5.4	7.0
	Other	618	4.86	1.80	1.0	9.0	3.6	4.9	5.9	7.0
Chemistry	UK	1,550	4.33	1.49	1.0	9.0	3.4	4.4	5.2	6.2
	EU	109	3.82	1.92	1.0	8.9	2.3	3.7	5.2	6.4
	Other	1,162	5.08	1.89	1.0	9.0	3.8	4.9	6.4	7.6
Physics	UK	4,657	4.15	1.44	1.0	9.0	3.2	4.1	5.0	6.0
	EU	651	3.85	1.51	1.0	8.9	2.8	3.8	4.8	5.9
	Other	3,929	5.03	1.80	1.0	9.0	3.8	4.9	6.2	7.5
Maths 2	UK	4,929	4.07	1.41	1.0	9.0	3.1	4.1	5.0	5.7
	EU	673	3.91	1.43	1.0	8.9	3.1	4.1	4.9	5.6
	Other	4,438	5.11	1.79	1.0	9.0	3.8	4.9	6.3	7.6

Figure 9 Box and Whisker Plot of Scaled Score 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, around half identified as British followed by 20% to 24% identifying as Chinese (Figure 10). The summary statistics for UK vs Chinese candidates are shown in Table 9. This shows that the candidates who identified as Chinese significantly outperformed UK nationals across all modules. The mean score difference ranged from 0.42 for Biology to 1.98 for Maths 1. The gap is largest for Maths 1 and Maths 2.

Figure 10 Percent of Candidates by Nationality

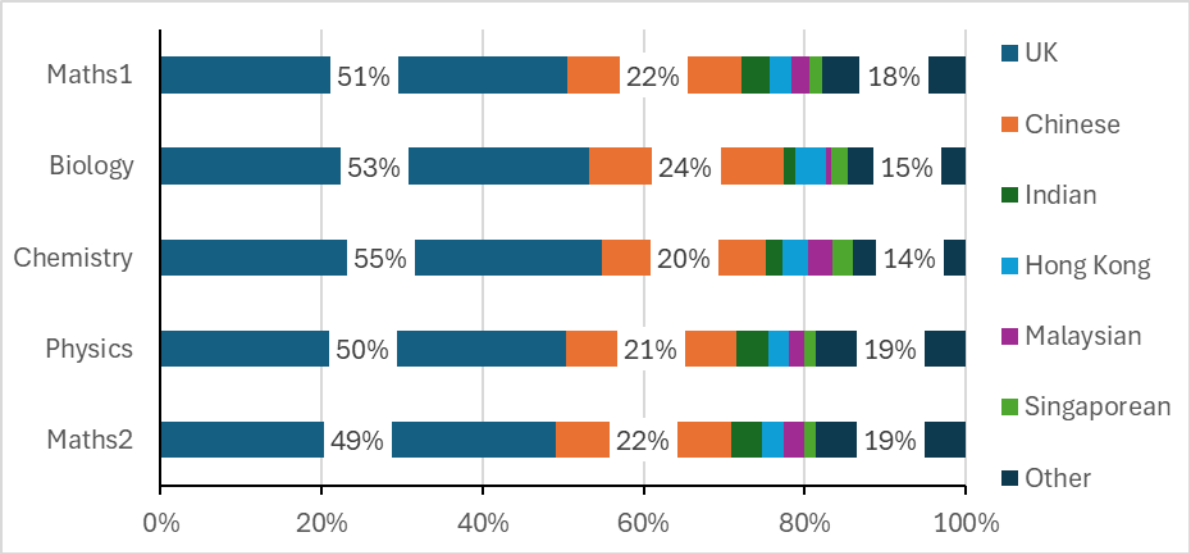


Table 9 Scaled Score Summary Statistics by Nationality

Module	Country	*N*	Scaled Score				Percentile			
			Mean	*SD*	Min	Max	25	50	75	90
Maths 1	UK	6,031	3.93	1.35	1.0	9.0	3.1	3.9	4.8	5.6
	Chinese	2,568	5.91	1.70	1.0	9.0	4.7	5.8	7.1	8.5
Biology	UK	762	4.64	1.67	1.0	9.0	3.6	4.5	5.4	7.0
	Chinese	345	5.06	1.68	1.0	9.0	4.0	5.0	6.4	7.6
Chemistry	UK	1,550	4.33	1.49	1.0	9.0	3.4	4.4	5.2	6.2
	Chinese	574	5.60	1.74	1.0	9.0	4.5	5.6	6.8	8.2
Physics	UK	4,657	4.15	1.44	1.0	9.0	3.2	4.1	5.0	6.0
	Chinese	1,961	5.58	1.75	1.0	9.0	4.5	5.6	6.8	8.0
Maths 2	UK	4,929	4.07	1.41	1.0	9.0	3.1	4.1	5.0	5.7
	Chinese	2,197	5.68	1.75	1.0	9.0	4.5	5.6	6.8	8.2

4.2.4 Ethnicity (UK Candidates Only)

UAT 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 11 shows the breakdown of candidates by ethnicity for the ESAT modules. The largest ethnic group amongst UK candidates was White (43% for Maths 2 to 67% for Biology) closely followed by Asian (18% for Biology to 35% for Maths 2). The candidates who chose to take Biology were the least diverse, with only 33% of the UK students being non-White. This is in contrast to Maths 2, where over half (57%) of UK candidates were non-White.

Figure 11 Percent of UK Candidates by Ethnicity

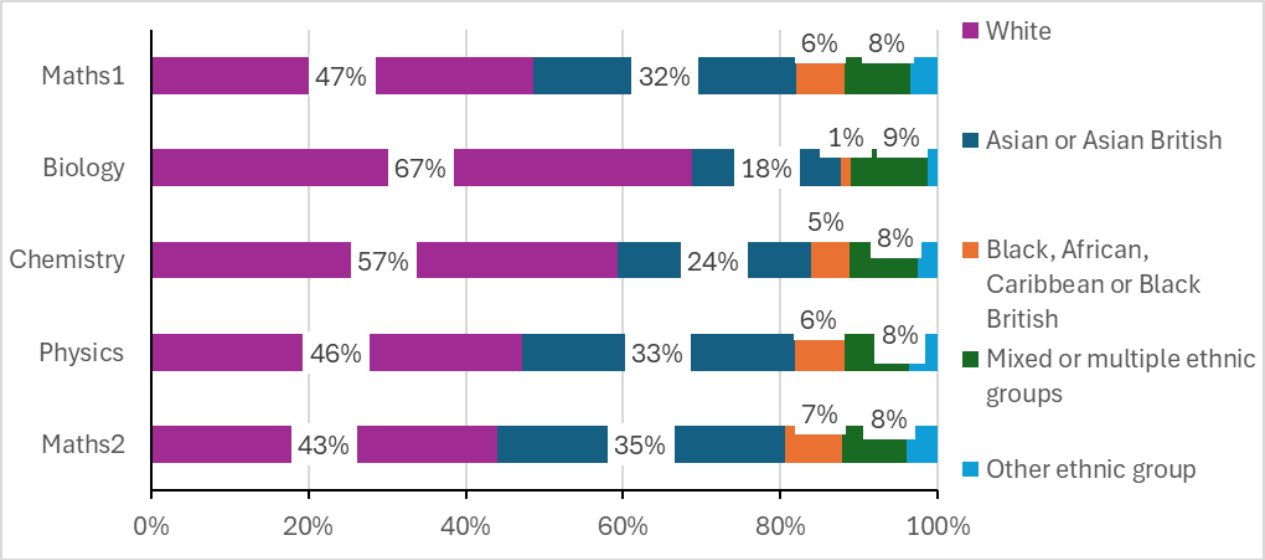


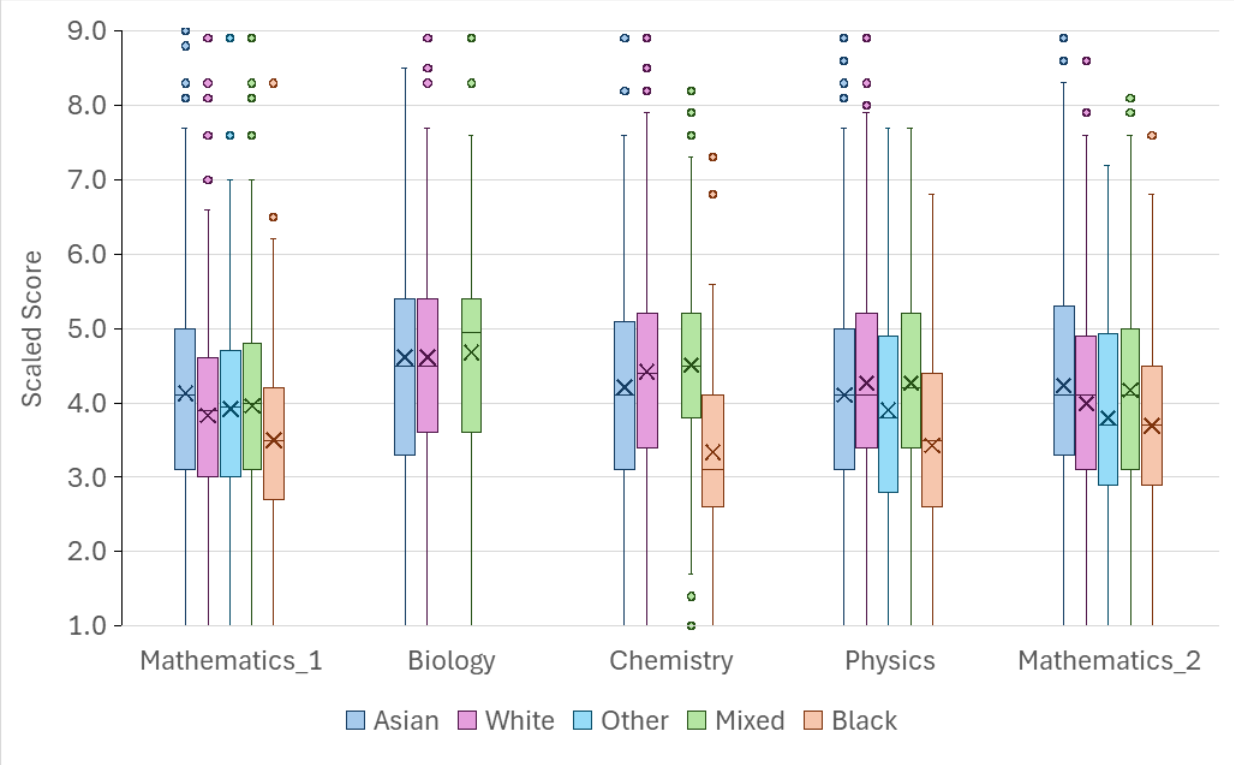
Table 10 summarises the scaled score summary statistics by ethnic group for UK students across the modules, for groups with at least 50 candidates. The differences in scaled score across the ethnic groups in the UK tend to reflect underlying trends within the UK.

Table 10 Scaled Score Summary Statistics by Ethnicity for UK Candidates

Module	Ethnicity	*N*	Scaled Score				Percentile			
			Mean	S*D*	Min	Max	25	50	75	90
Maths 1	Asian	1,942	4.13	1.46	1.0	9.0	3.1	4.1	5.0	5.9
	Black	363	3.50	1.23	1.0	8.3	2.7	3.5	4.2	5.1
	Mixed	483	3.96	1.31	1.0	8.9	3.1	4.0	4.8	5.4
	Other	208	3.92	1.42	1.0	8.9	3.0	4.0	4.7	5.6
	White	2,837	3.83	1.26	1.0	8.9	3.0	3.9	4.5	5.4
Biology	Asian	140	4.62	1.78	1.0	9.0	3.4	4.5	5.4	7.3
	Black	9	NA	NA	NA	NA	NA	NA	NA	NA
	Mixed	72	4.68	1.81	1.0	9.0	3.6	5.0	5.4	7.0
	Other	10	NA	NA	NA	NA	NA	NA	NA	NA
	White	509	4.62	1.61	1.0	9.0	3.6	4.5	5.4	7.0
Chemistry	Asian	369	4.21	1.51	1.0	8.9	3.1	4.1	5.1	6.4
	Black	73	3.34	1.22	1.0	7.3	2.6	3.1	4.1	4.8
	Mixed	129	4.51	1.48	1.0	8.2	3.8	4.5	5.2	6.4
	Other	40	NA	NA	NA	NA	NA	NA	NA	NA
	White	889	4.42	1.46	1.0	9.0	3.4	4.4	5.2	6.2
Physics	Asian	1,560	4.11	1.47	1.0	9.0	3.1	4.1	5.0	6.0
	Black	285	3.43	1.28	1.0	6.8	2.7	3.5	4.4	4.9
	Mixed	369	4.27	1.52	1.0	8.3	3.4	4.2	5.2	6.3
	Other	167	3.90	1.54	1.0	9.0	2.8	3.8	4.9	6.0
	White	2,119	4.27	1.38	1.0	9.0	3.4	4.1	5.2	6.0
Maths 2	Asian	1,746	4.24	1.45	1.0	9.0	3.3	4.1	5.3	6.0
	Black	343	3.69	1.28	1.0	7.6	2.9	3.7	4.5	5.0
	Mixed	389	4.17	1.43	1.0	8.1	3.1	4.1	5.0	6.1
	Other	194	3.80	1.37	1.0	7.2	2.9	3.7	4.9	5.4
	White	2,098	4.00	1.36	1.0	8.6	3.1	4.1	4.9	5.6

Figure 12 shows a box and whisker plot of scaled scores by ethnicity for each module. For Maths 1 and Maths 2, the strongest UK group was Asian with a mean scaled score of 4.13 and 4.24, respectively. For Biology, Chemistry and Physics the strongest UK ethnic group was Mixed. The poorest performing UK ethnic group was those identifying as Black. This is in line with national trends.

Figure 12 Box and Whisker Plot of Scaled Score by Ethnicity



4.2.5 First Language

Figure 13 illustrates the proportion of candidates who identified English as their first language by module. Maths 2 had the highest proportion of candidates with English as a first language (65%), and Physics and Maths 1 had the lowest of 56%. Interestingly, candidates with English as a first language were weaker than those whose first language was not English (Table 11) across all modules.

Figure 13 Percent of Candidates by First Language

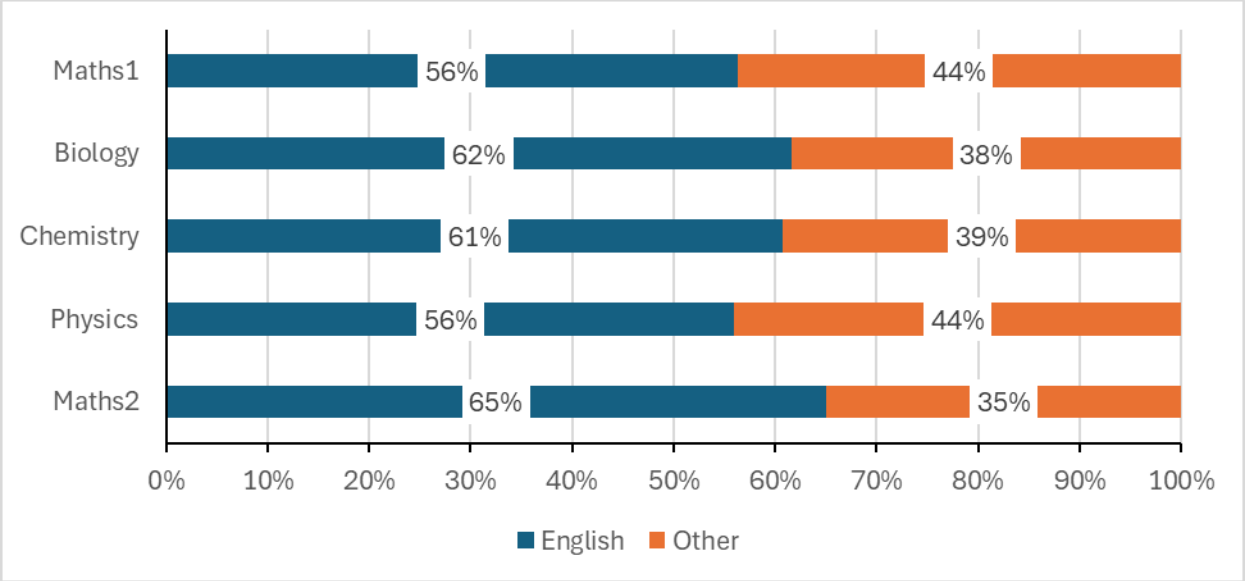
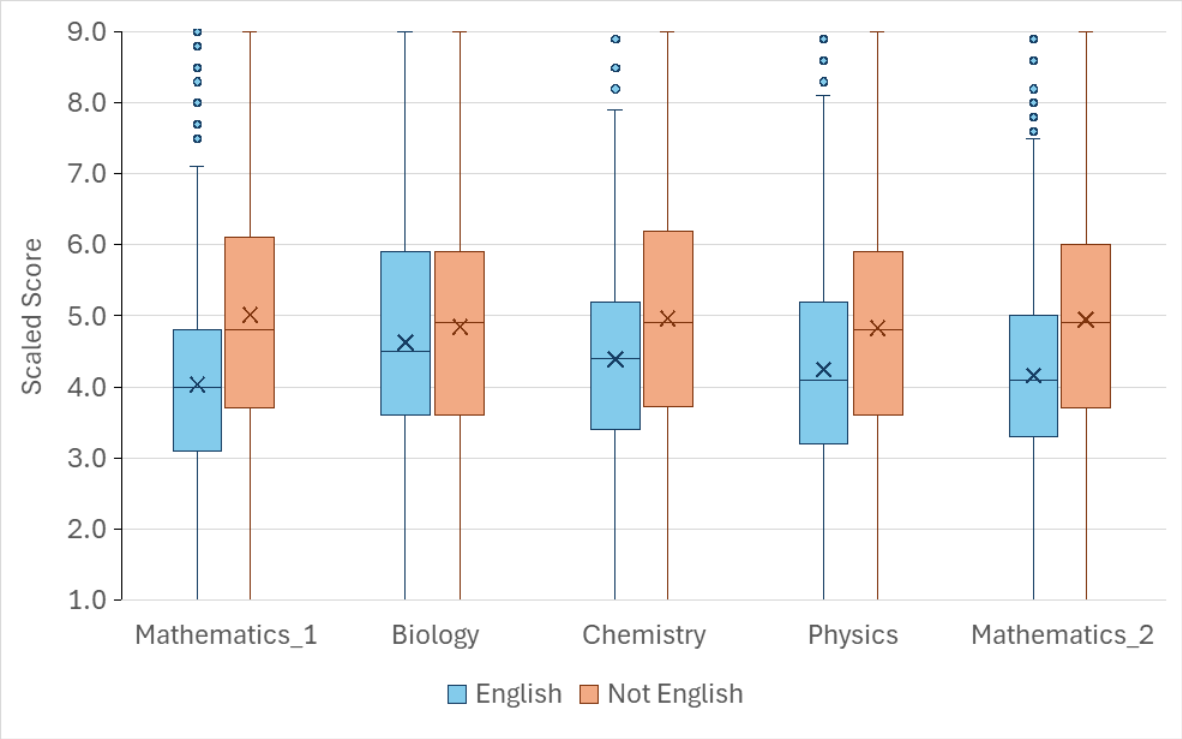


Table 11 Scaled Score Summary Statistics by First Language

Module	First Language	*N*	Scaled Score				Percentile			
			Mean	*SD*	Min	Max	25	50	75	90
Maths 1	English	6,709	4.03	1.43	1.0	9.0	3.1	4.0	4.8	5.8
	Other	5,210	5.01	1.87	1.0	9.0	3.7	4.8	6.1	7.6
Biology	English	881	4.63	1.71	1.0	9.0	3.6	4.5	5.9	7.0
	Other	548	4.85	1.79	1.0	9.0	3.6	4.9	5.9	7.0
Chemistry	English	1,713	4.39	1.58	1.0	9.0	3.4	4.4	5.2	6.4
	Other	1,108	4.96	1.89	1.0	9.0	3.8	4.9	6.2	7.5
Physics	English	5,173	4.25	1.49	1.0	9.0	3.2	4.1	5.2	6.3
	Other	4,064	4.83	1.82	1.0	9.0	3.6	4.8	5.9	7.2
Maths 2	English	5,473	4.16	1.45	1.0	9.0	3.3	4.1	5.0	6.0
	Other	4,567	4.95	1.82	1.0	9.0	3.7	4.9	6.0	7.6

Figure 14 shows a box and whisker plot of scaled scores by first language for each module. This illustrates that for all modules, candidates who did not speak English as a first language outperformed those who did. The gap is smallest for Biology, where the mean score for non-English first language candidates was 0.22 higher than those with English as a first language, and largest for Maths 1 (0.98). This result is primarily due to a very strong cohort in China but is especially interesting for those subjects with a higher reading load such as Biology; therefore, it is expected that there is a smaller difference in ability between cohorts in Biology.

Figure 14 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 across all modules, ranging from 38% (Chemistry and Biology) to 42% (Maths 2) of candidates, was a Further Education College or Sixth Form College (Figure 15). The least common school type in the UK was an Academy/Secondary School, with 16% to 19% of candidates selecting these school types across the modules. Grammar school candidates made up 17% to 20% of the UK candidates taking ESAT. This is much higher than the national average (around 5% of state-funded secondary pupils attend grammar schools).

The scaled score summary statistics for groups with at least 50 candidates can be found in Table 12. Figure 16 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, and Grammar and Private/Fee Paying Schools had higher mean scaled scores. This is in line with national trends, where selective and private schools tend to outperform other types of schools.

Figure 15 Percent of Candidates by School Type

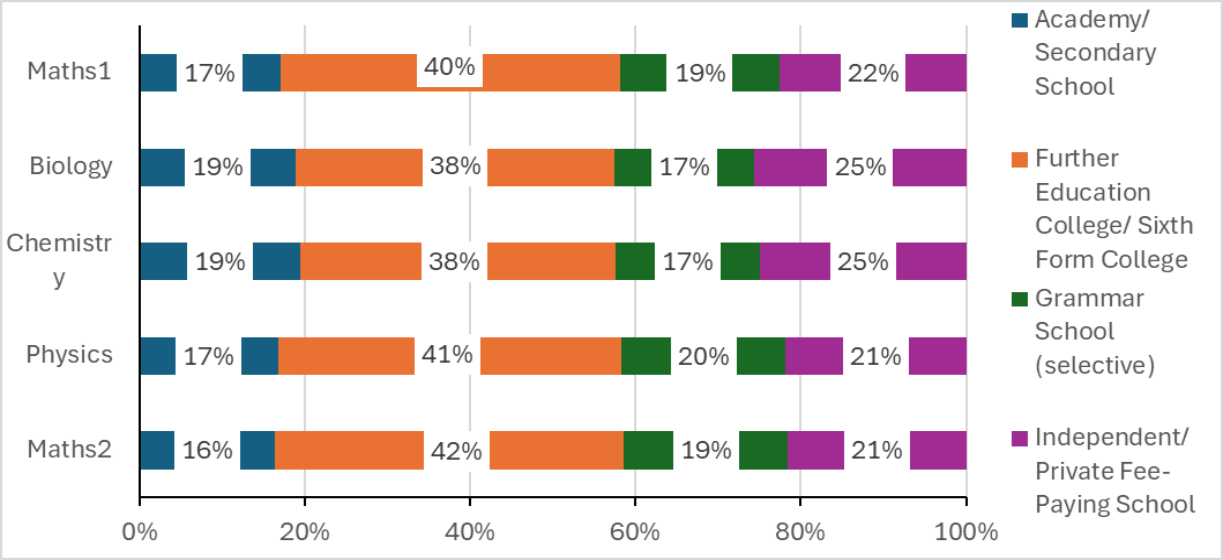
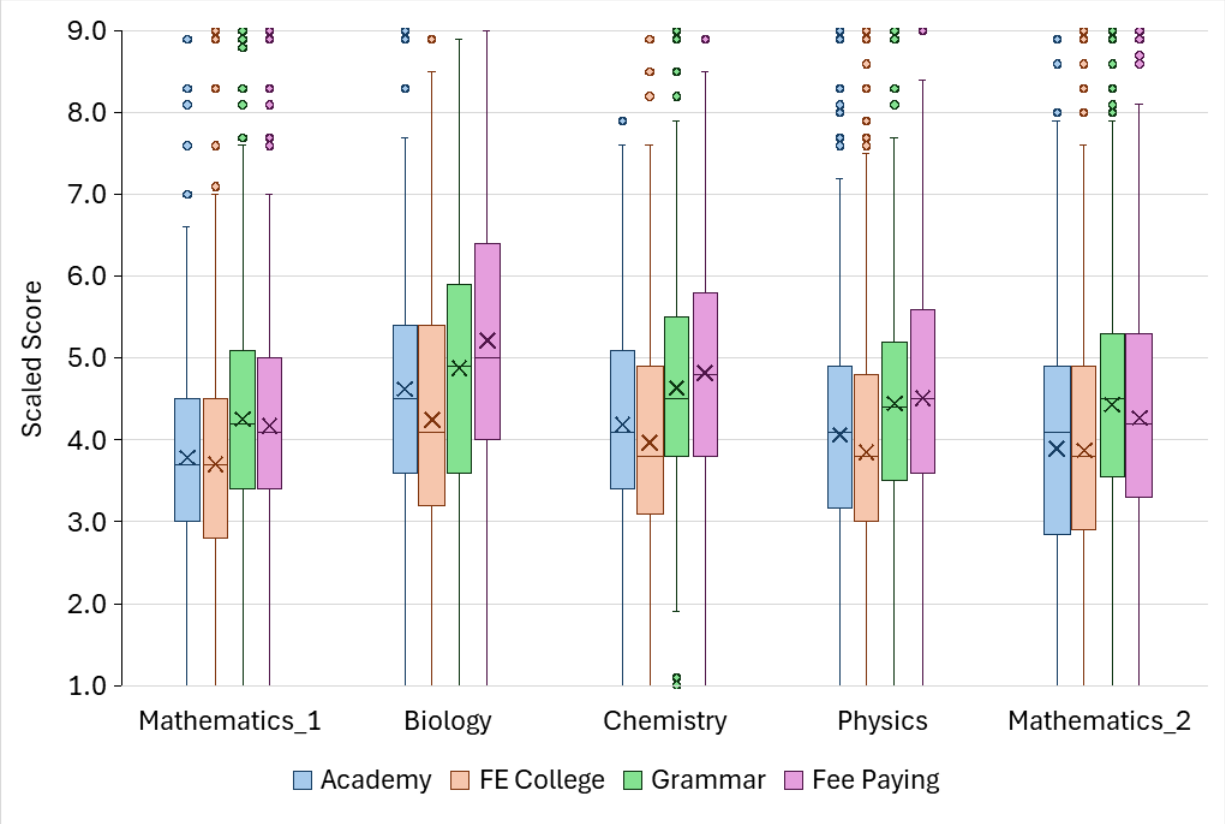


Table 12 Scaled Score Summary Statistics by School Type

Module	School Type	*N*	Scaled Score				Percentile			
			Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Secondary School	1,011	3.79	1.31	1.0	8.9	3.0	3.7	4.5	5.4
	FE/ 6th Form College	2,437	3.70	1.29	1.0	9.0	2.8	3.7	4.5	5.4
	Grammar School	1,147	4.26	1.36	1.0	9.0	3.4	4.2	5.1	5.9
	Private Fee-Paying School	1,339	4.18	1.38	1.0	9.0	3.4	4.1	5.0	5.8
	University	59	3.53	1.55	1.0	7.0	2.7	3.6	4.7	5.8
Biology	Secondary School	141	4.62	1.59	1.0	9.0	3.6	4.5	5.4	7.0
	FE/ 6th Form College	287	4.25	1.53	1.0	8.9	3.2	4.1	5.4	5.9
	Grammar School	126	4.88	1.63	1.0	8.9	3.6	4.9	5.9	7.0
	Private Fee-Paying School	191	5.22	1.68	1.0	9.0	4.0	5.0	6.4	7.6
	University	12	NA	NA	NA	NA	NA	NA	NA	NA
Chemistry	Secondary School	297	4.19	1.32	1.0	7.9	3.4	4.1	5.1	6.0
	FE/ 6th Form College	582	3.97	1.43	1.0	8.9	3.1	3.8	4.9	5.8
	Grammar School	267	4.64	1.52	1.0	9.0	3.8	4.5	5.5	6.8
	Private Fee-Paying School	380	4.82	1.50	1.0	8.9	3.8	4.8	5.8	6.6
	University	15	NA	NA	NA	NA	NA	NA	NA	NA
Physics	Secondary School	770	4.07	1.34	1.0	9.0	3.2	4.1	4.9	5.9
	FE/ 6th Form College	1,904	3.85	1.40	1.0	9.0	3.0	3.8	4.8	5.6
	Grammar School	912	4.45	1.44	1.0	9.0	3.5	4.4	5.2	6.3
	Private Fee-Paying School	1,001	4.51	1.44	1.0	9.0	3.6	4.5	5.6	6.3
	University	41	NA	NA	NA	NA	NA	NA	NA	NA
Maths 2	Secondary School	797	3.88	1.33	1.0	9.0	2.9	3.8	4.9	5.6
	FE/ 6th Form College	2,048	4.43	1.43	1.0	9.0	3.7	4.5	5.3	6.3
	Grammar School	961	4.27	1.43	1.0	9.0	3.3	4.2	5.3	6.0
	Private Fee-Paying School	1,047	4.06	1.32	1.1	7.0	3.1	3.8	5.3	5.6
	University	45	NA	NA	NA	NA	NA	NA	NA	NA

Figure 16 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. Between 8% (Biology) and 11% (Maths 2) of the candidates who took ESAT received free school meals (Figure 17). This is lower than the almost 25% of pupils nationally who are eligible for free school meals. The scaled score summary statistics for these two groups of candidates are in Table 13 and are plotted in a box and whisker plot in Figure 18. This shows that the candidates who are or were receiving free school meals tend to have much lower scores than those who are or were not, which is in line with national trends.

Figure 17 Percent of Candidates by Free School Meals (UK Only)

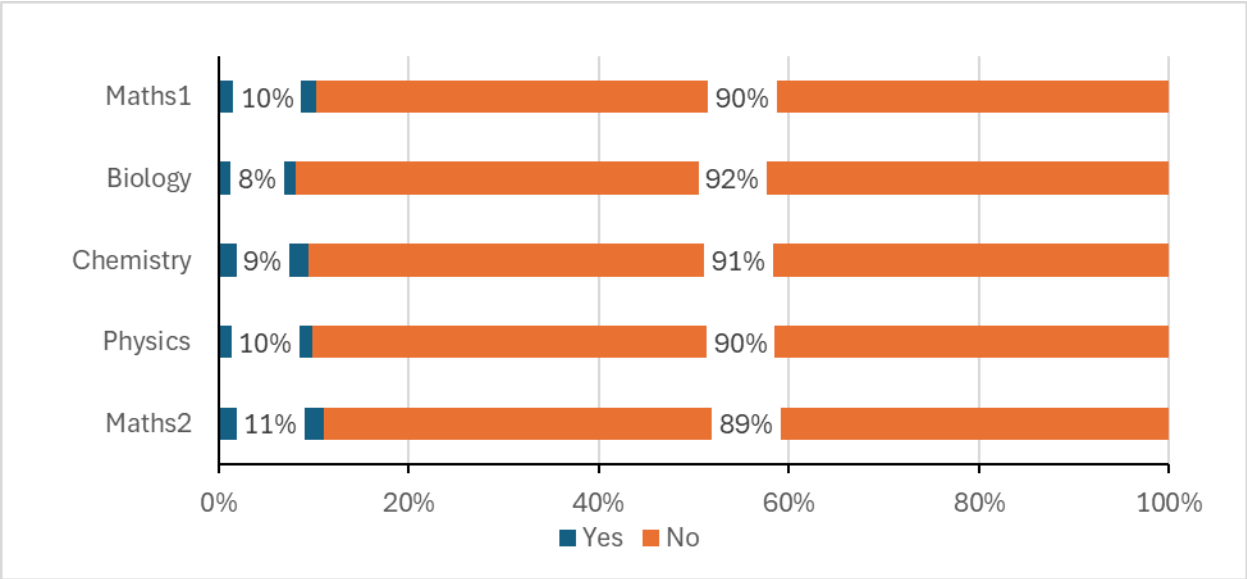
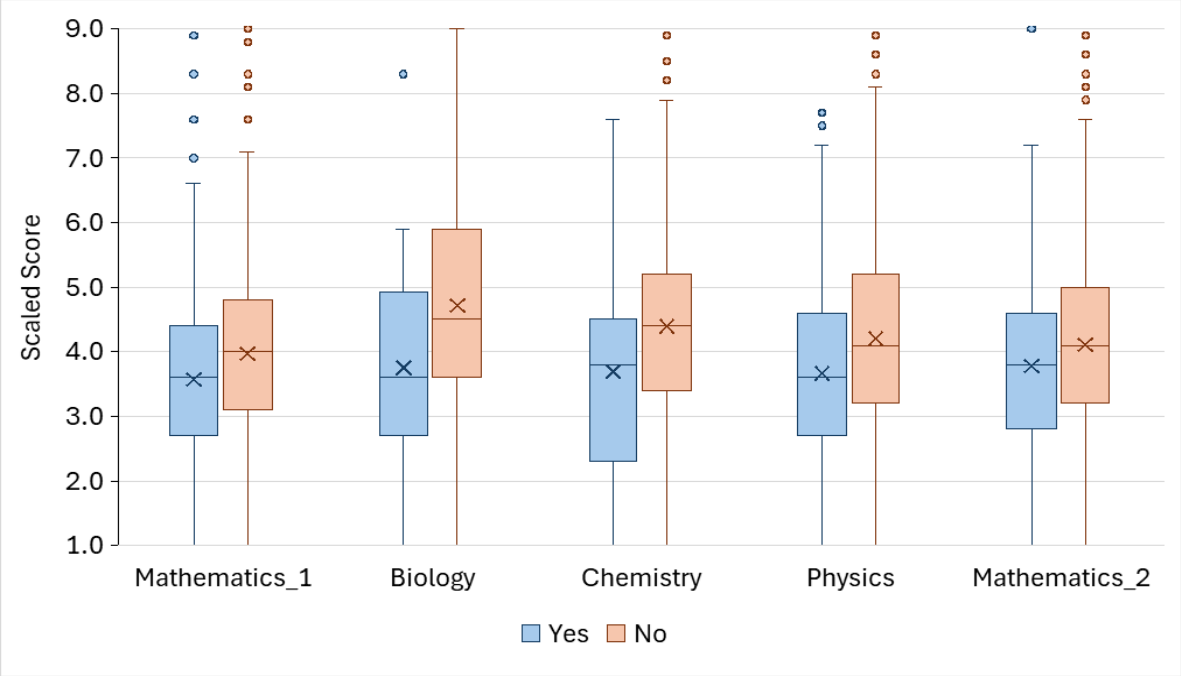


Table 13 Summary Statistics by Free School Meals (UK Only)

Module	Free School Meals	*N*	Scaled Score				Percentile			
			Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Yes	617	3.57	1.30	1.0	9.0	2.7	3.6	4.4	5.1
	No	5,414	3.97	1.35	1.0	9.0	3.1	4.0	4.8	5.8
Biology	Yes	62	3.75	1.69	1.0	8.3	2.7	3.6	4.9	5.9
	No	700	4.72	1.64	1.0	9.0	3.6	4.5	5.9	7.0
Chemistry	Yes	146	3.69	1.49	1.0	7.6	2.3	3.8	4.5	5.6
	No	1,404	4.39	1.48	1.0	9.0	3.4	4.4	5.2	6.4
Physics	Yes	462	3.67	1.40	1.0	7.7	2.7	3.6	4.6	5.2
	No	4,195	4.20	1.43	1.0	9.0	3.2	4.1	5.2	6.0
Maths 2	Yes	544	3.78	1.37	1.0	9.0	2.8	3.8	4.6	5.6
	No	4,385	4.11	1.41	1.0	9.0	3.2	4.1	5.0	6.0

Figure 18 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). Between 79% (Maths 1, Maths 2) and 82% (Biology) of candidates responded "Yes" (Figure 19). Responding "No" to this question can be considered, along with free school meals, as an indicator of widening participation of candidates. The scaled score summary statistics are shown in Table 14 and illustrated in a box and whisker plot in Figure 20. These data show that across all modules, the candidates who had a parent/guardian attend higher education significantly outperformed those who did not.

Figure 19 Percent of Candidates by Parent Higher Education

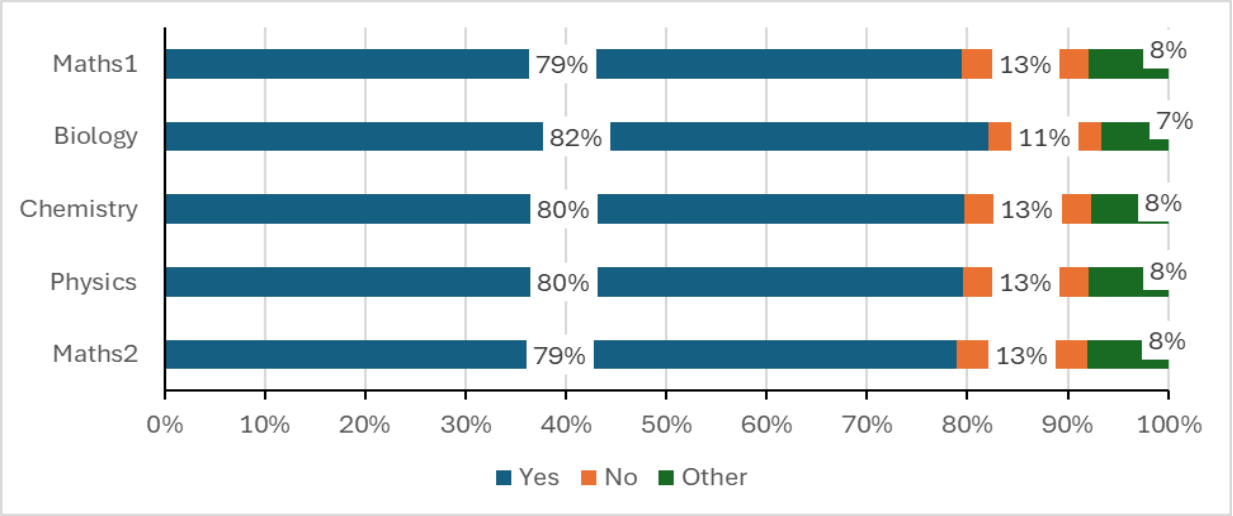
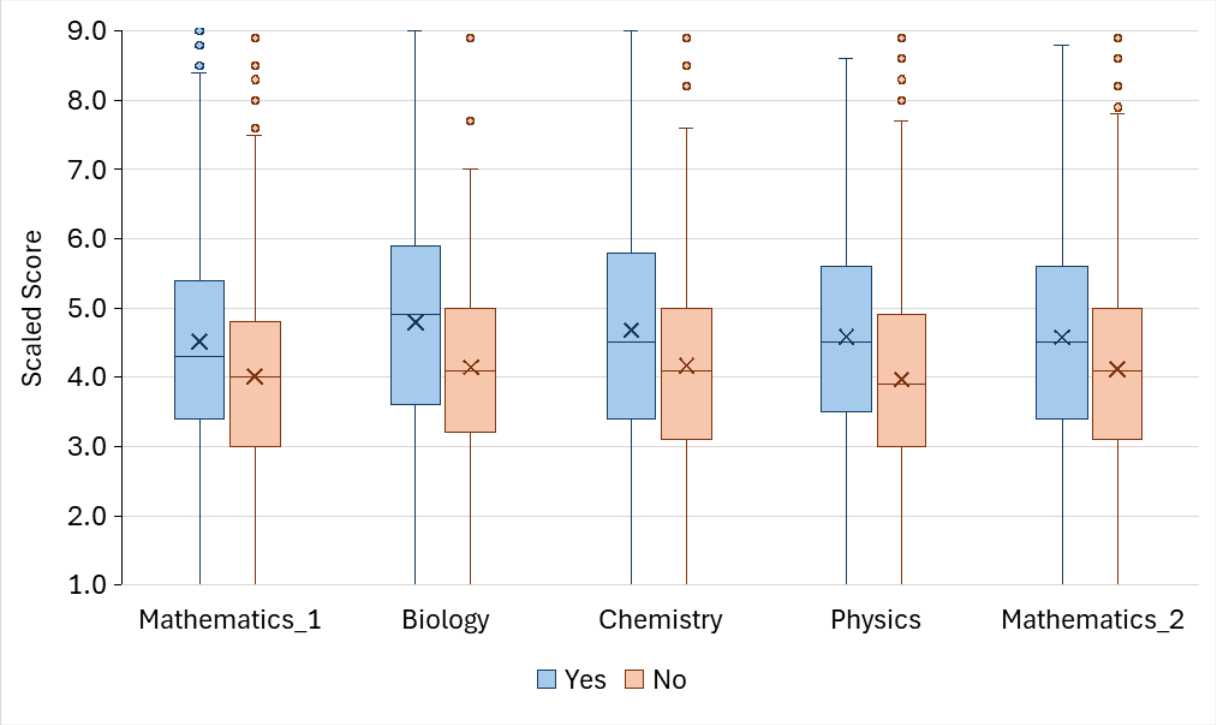


Table 14 Summary Statistics by Parent/Guardian Education

Module	Parent Higher Education	*N*	Scaled Score				Percentile			
			Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Yes	9,470	4.52	1.72	1.0	9.0	3.4	4.3	5.4	7.0
	No	1,506	4.02	1.55	1.0	9.0	3.0	4.0	4.8	6.0
Biology	Yes	1,173	4.80	1.78	1.0	9.0	3.6	4.9	5.9	7.0
	No	161	4.14	1.56	1.0	8.9	3.2	4.1	5.0	5.9
Chemistry	Yes	2,247	4.68	1.74	1.0	9.0	3.4	4.5	5.8	7.0
	No	359	4.17	1.62	1.0	9.0	3.1	4.1	5.0	6.4
Physics	Yes	7,349	4.58	1.68	1.0	9.0	3.5	4.5	5.6	6.8
	No	1,157	3.97	1.50	1.0	9.0	3.0	3.9	4.9	5.9
Maths 2	Yes	7,922	4.57	1.69	1.0	9.0	3.4	4.5	5.6	6.8
	No	1,304	4.12	1.53	1.0	9.0	3.1	4.1	5.0	6.0

Figure 20 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. Across the five modules, 89% to 90% of candidates said that they did not have such a health condition or disability (Figure 21). There is a difference in performance across the two groups, with those who responded "No" outperforming those who responded "Yes" (Table 15; Figure 22).

Figure 21 Percent of Candidates by Learning Difficulty/Chronic Health Condition

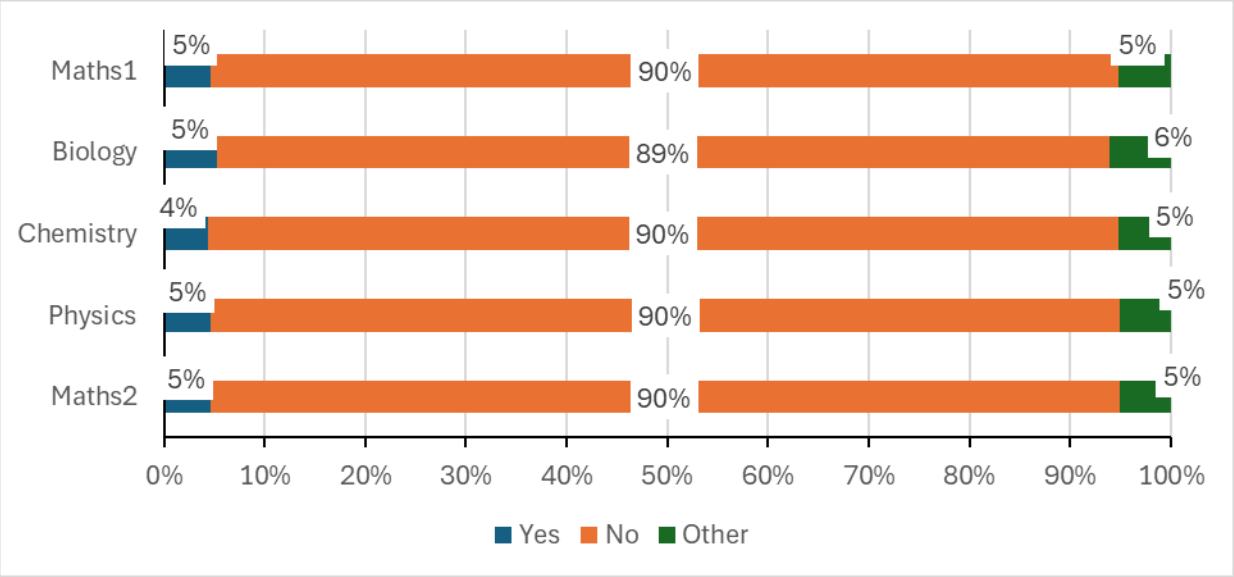
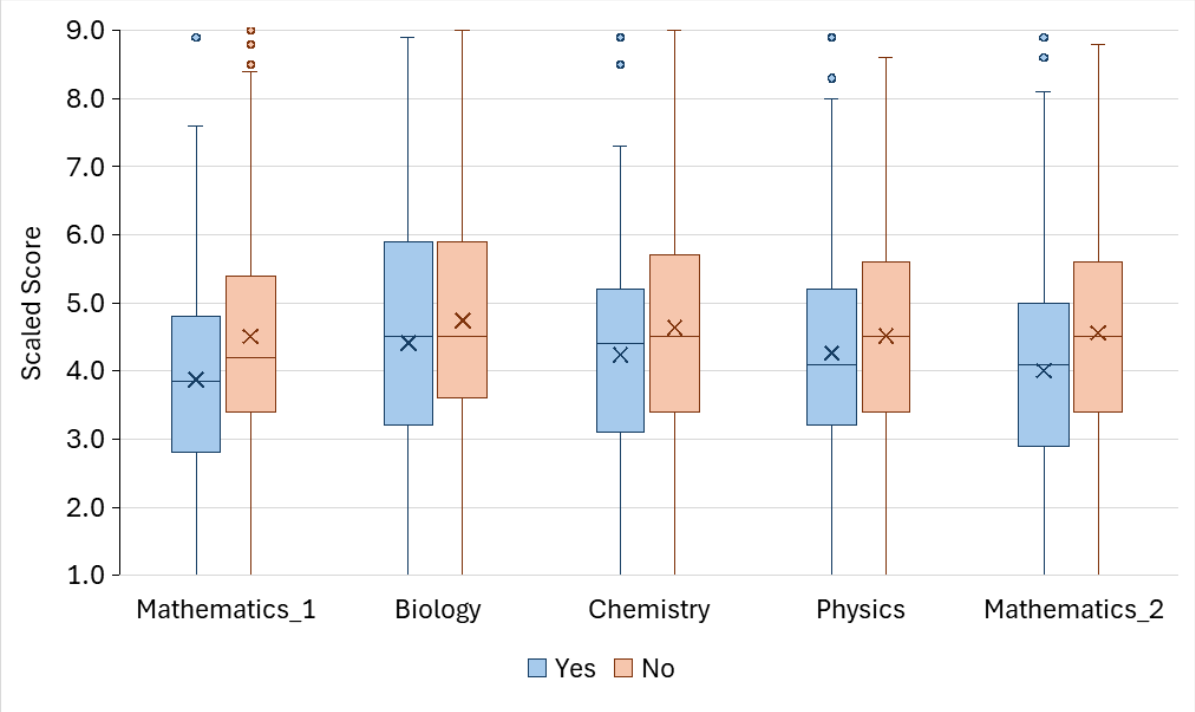


Table 15 Summary Statistics by Learning Difficulty/Chronic Health Condition

Module	Learning Difficulty	*N*	Scaled Score				Percentile			
			Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Yes	558	3.87	1.46	1.0	9.0	2.8	3.9	4.8	5.6
	No	10,739	4.51	1.72	1.0	9.0	3.4	4.2	5.4	7.0
Biology	Yes	75	4.41	1.74	1.0	8.9	3.2	4.5	5.9	6.4
	No	1,266	4.74	1.75	1.0	9.0	3.6	4.5	5.9	7.0
Chemistry	Yes	123	4.24	1.69	1.0	8.9	3.1	4.4	5.2	6.5
	No	2,550	4.64	1.73	1.0	9.0	3.4	4.5	5.7	7.0
Physics	Yes	430	4.27	1.52	1.0	9.0	3.2	4.1	5.2	6.3
	No	8,338	4.52	1.68	1.0	9.0	3.4	4.5	5.6	6.8
Maths 2	Yes	461	4.00	1.58	1.0	9.0	2.9	4.1	5.0	5.7
	No	9,062	4.56	1.68	1.0	9.0	3.4	4.5	5.6	6.8

Figure 22 Box and Whisker Plot of Scaled Score by Learning Disability/Chronic Health Condition



4.3 Accommodations

Candidates can request accommodations, such as extra time, if required for their test. In total, 628 candidates had accommodations for Maths 1, the compulsory module, which is around 5% of the candidate population. Candidates can apply for more than one accommodation. Table 16 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 439 candidates (70% of Maths 1 accommodations candidates) requesting this.

Table 16 Number of Accommodations by Type

Accommodation	Maths 1	Biology	Chemistry	Physics	Maths 2
Extra Time	267	22	52	217	226
Pause-the-clock	43	7	11	35	32
Extra time + pause-the-clock	129	17	24	101	108
Separate Room	111	18	28	85	84
Other	78	14	17	57	62

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 ESAT modules each have only 27 items, making them relatively short modules. The reliabilities are therefore expected to be lower for ESAT than for TMUA as reliability is closely related to test length. A reliability of over 0.70 would still be considered satisfactory for the lengths of the ESAT modules.

The raw score reliabilities for Mathematics 1 were excellent, ranging from 0.74 to 0.83. Maths 1 performed the strongest in terms of internal consistency. For Biology there was a much smaller candidate sample and the reliabilities ranged from 0.63 to 0.74. For Chemistry, reliabilities for the October forms were excellent and ranged from 0.75 to 0.78. For the January forms, there was a much smaller number of candidates, and therefore these values may not be representative. The forms for Physics and Mathematics 2 were too difficult for many candidates, resulting in (effectively) shorter tests, and the maximum raw scores in some cases were below the full mark. This impacted reliabilities, which were slightly lower for these modules, ranging from 0.64 to 0.79 for Physics and 0.54 to 0.77 for Maths 2. Improving these modules by better targeting candidates' ability could enhance reliability. Reliability in all modules was generally better in October as candidates were stronger and therefore slightly better matched to the test difficulty.

Table 17 Raw Score Reliability and SEM

Module	Event	Raw Score Reliability	
		Cronbach's Alpha	*SEM*
Maths 1	Oct 2024	0.78 to 0.82	1.94 to 2.15
	Jan 2025	0.74 to 0.77	2.10 to 2.19
Biology	Oct 2024	0.63 to 0.74	2.13 to 2.14
	Jan 2025	-	-
Chemistry	Oct 2024	0.75 to 0.78	2.10 to 2.21
	Jan 2025	0.65 to 0.79	2.20 to 2.24
Physics	Oct 2024	0.65 to 0.79	2.06 to 2.21
	Jan 2025	0.66 to 0.73	2.15 to 2.22
Maths 2	Oct 2024	0.56 to 0.77	2.10 to 2.20
	Jan 2025	0.53 to 0.65	02.17 to 02.24

As ESAT candidates also receive for each module 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 18. This shows that the scaled score reliabilities are very similar to the raw score reliabilities, with Maths 1 having the highest reliabilities and Maths 2 having the lowest reliabilities.

Table 18 Scaled Score Reliability and SEM

Module	Event	Scaled Score Reliability	
		Reliability	*SEM*
Maths 1	Oct 2024	0.79 to 0.82	0.69 to 0.75
	Jan 2025	0.75 to 0.78	0.67 to 0.68
Biology	Oct 2024	0.65 to 0.74	0.96 to 0.97
	Jan 2025	-	-
Chemistry	Oct 2024	0.76 to 0.77	0.81 to 0.88
	Jan 2025	0.66 to 0.79	0.78 to 0.83
Physics	Oct 2024	0.67 to 0.79	0.83 to 0.84
	Jan 2025	0.64 to 0.72	0.81 to 0.86
Maths 2	Oct 2024	0.54 to 0.77	0.81 to 0.96
	Jan 2025	0.54 to 0.65	0.90 to 0.91

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 19. Candidates with a time over 00:40:04— approximately 3% per module—were excluded from analysis of module time as they are assumed to have an accommodation (the number of these candidates is listed in Table 20). The mean module time for each of the modules is between 39 and 40 minutes. The median module time is a better indication of average module time as extreme times have a smaller impact on the median than the mean. The median is 40 minutes for all modules except Biology, which is 39 minutes and 53 seconds. The distribution of candidates by module time is illustrated in Figure 23, which shows that over 90% of candidates used between 35 and 40 minutes.

Table 19. Test Time Summary Statistics

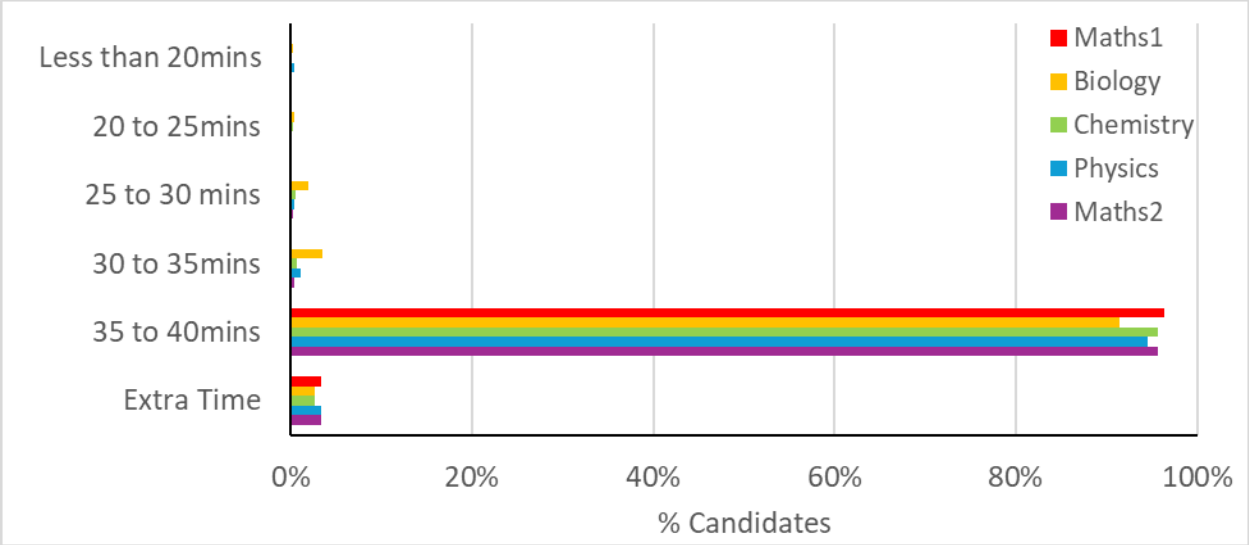
Module	*N*	Test Time				
		Mean	Median	*SD*	Min	Max
Maths 1	11,523	00:39:51	00:40:00	00:00:38	00:12:01	00:40:02
Biology	1,391	00:39:00	00:39:53	00:02:42	00:04:09	00:40:01
Chemistry	2745	00:39:36	00:40:00	00:01:47	00:05:30	00:40:02
Physics	8925	00:39:31	00:40:00	00:02:28	00:00:00	00:40:02
Maths 2	9708	00:39:42	00:40:00	00:01:35	00:00:00	00:40:02

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

Table 20 Candidates with Extra Time

Module	Test Time Greater Than 00:40:04	
	N	%
Maths 1	396	3%
Biology	38	3%
Chemistry	76	3%
Physics	312	3%
Maths 2	332	3%

Figure 23 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 could be speeded. Speededness can be further assessed by looking at the number of unreached, or not presented, items. The ESAT 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 21. This shows that for Biology (3% of candidates) and Chemistry (5% of candidates), only a very small proportion of candidates had unreached items. This is what would be expected for a test of this type, so it is likely that these modules are not speeded. There is a moderate degree of speeding in Physics, where 10% of candidates had at least one unreached item and the mean number of unreached items was 3.22. The most speeded modules were Maths 1 and Maths 2, where 15% and 17% of candidates, respectively, had at least one unreached item. It is likely that these two modules are speeded.

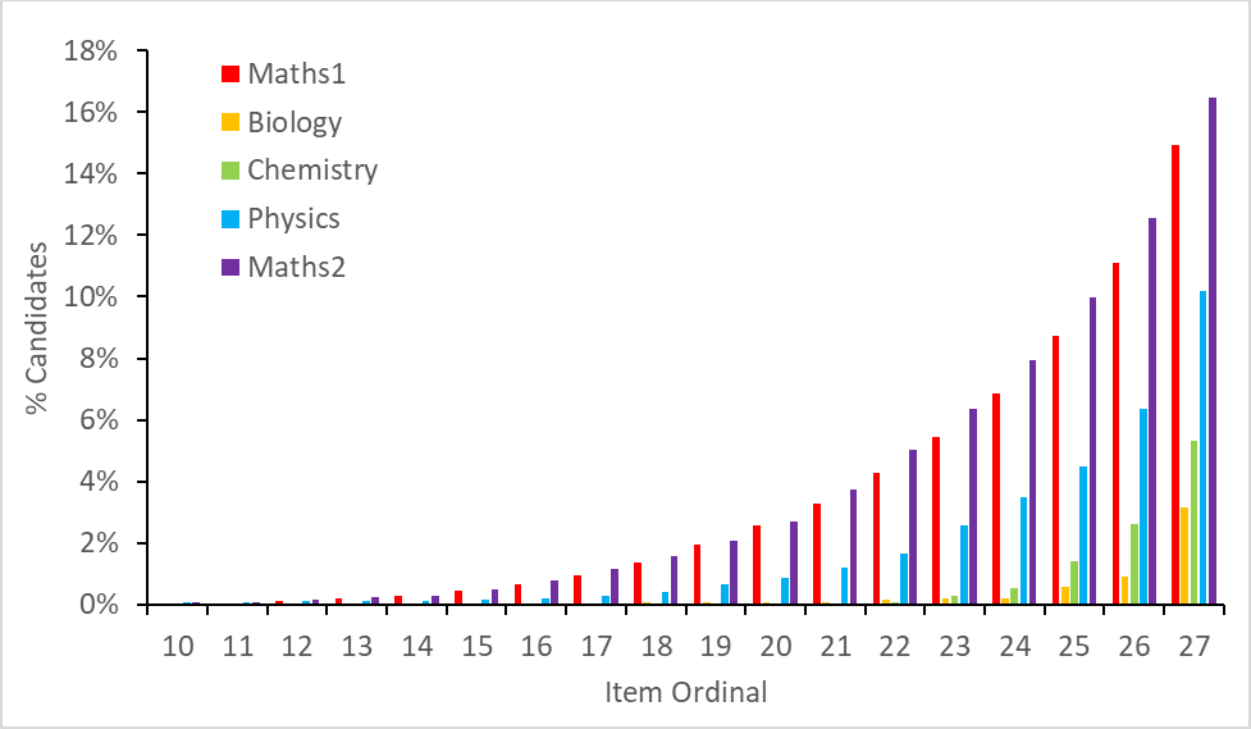
To reduce speededness, item time could be considered when building the forms in the future as long as the item bank is large enough to support this. Item time statistics should also be reviewed to identify any trends amongst items that are taking more time.

Table 21 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
Maths 1	11,919	1,813	15%	4.16	1 to 18
Biology	1,429	46	3%	1.70	1 to 9
Chemistry	2,821	150	5%	1.95	1 to 9
Physics	9,237	956	10%	3.22	1 to 27
Maths 2	10,040	1,677	17%	4.31	1 to 27

The percentage of candidates with unreached items by item ordinal is plotted in Figure 24 (from Item 10). This plot includes all unreached items, so if a candidates did not reach items 25, 26 and 27, 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 24 shows that Maths 1 and Maths 2 are the most speeded and Biology and Chemistry are the least.

Figure 24 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 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 23 summarises the number of candidates with unreached or low time items. The trends observed very much mirror those observed in Table 22, with the percentage of candidates with low

time/unreached items lowest for Biology (6%) and Chemistry (15%) and the mean number of low time/unreached items being 2.00 or less for these two modules. The most speeded module is shown to be Maths 2, with over 50% of candidates having at least one low time/unreached item and a mean number of low time/unreached items of 3.90. This is in line with other data for Maths 2 that showed that overall the module was too difficult for the candidates. Reviewing the items with an aim of reducing the mean time per item could contribute to lowering the overall difficulty of the Maths 2 module.

Table 23 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
Maths 1	11,919	5,440	46%	3.60	1 to 20
Biology	1,429	80	6%	1.92	1 to 16
Chemistry	2,821	414	15%	2.00	1 to 16
Physics	9,237	2,824	31%	3.08	1 to 27
Maths 2	10,040	5,230	52%	3.90	1 to 27

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 ESAT Item Analysis

For all ESAT modules, 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.

The outcome of the item analyses for each of the ESAT modules are summarised in Tables 24 to 28.

6.1.1 Mathematics 1 Item Performance

Of the items used in the Mathematics 1 forms, 99% of items met the criteria — well above the 80% target for new exams. All items had a positive point biserial, ranging from 0.16 to 0.59 with a mean of 0.38, which is excellent.

Table 24 Mathematics 1 Item Status Outcome

Status	Comment	% of Items
Fail	*b* Value greater than +3 (difficult item)	1%
	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	6%
	None	89%
Total		100%

6.1.2 Biology Item Performance

Of the items used in the Biology forms, 98% of items met the criteria — well above the 80% target for new exams. All items had a positive point biserial ranging from 0.19 to 0.49 with a mean of 0.33, which is good.

Table 25 Biology Item Status Outcome

Status	Comment	% of Items
Fail	*b* Value greater than +3 (difficult item)	0%
	b Value less than -3 (easy item)	2%
	Very low or negative point biserial	0%
Pass	*p* Value greater than 0.90	4%
	p Value less than 0.20	2%
	None	92%
Total		100%

6.1.3 Chemistry Item Performance

Of the items used in the Chemistry forms, 99% of items met the criteria — well above the 80% target for new exams. All items had a positive point biserial ranging from 0.16 to 0.54 with a mean of 0.37, which is good.

Table 26 Chemistry Item Status Outcome

Status	Comment	% of Items
Fail	*b* Value greater than +3 (difficult item)	0%
	b Value less than -3 (easy item)	1%
	Very low or negative point biserial	0%
Pass	*p* Value greater than 0.90	3%
	p Value less than 0.20	1%
	None	94%
Total		100%

6.1.4 Physics Item Performance

Of the items used in the Physics forms, 99% of items met the criteria — well above the 80% target for new exams. All items had a positive point biserial ranging from 0.11 to 0.51 with a mean of 0.35, which is good.

Table 27 Physics Item Status Outcome

Status	Comment	% of Items
Fail	*b* Value greater than +3 (difficult item)	1%
	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	9%
	None	87%
Total		100%

6.1.5 Mathematics 2 Item Performance

Of the items used in the Mathematics 2 forms, 98% of items met the criteria — well above the 80% target for new exams — with 2% failing to meet the criteria because of their low point biserial. However, all items had a positive point biserial, ranging from 0.04 to 0.54 with a mean of 0.31, which is good. It should be noted that 16% of the items had a *p* value below 0.20, which indicates that these items are too difficult for the majority of the cohort.

Table 28 Mathematics 2 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	2%
Pass	*p* Value greater than 0.90	0%
	p Value less than 0.20	16%
	None	84%
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 ESAT 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

Mathematics 1

The DIF results are reported in Table 29 for Maths 1 items that showed Category C DIF. These are items where there is a significant difference in the performance of candidates in different demographic groups.

Four items were flagged as showing DIF by first language, with one item favouring candidates spoke English as a first language and three favouring candidates who did not speak English as their first language. These items are across a range of item difficulties, although three of the four items have more extreme difficulties (two difficult and one easy). These items will be reviewed to identify likely sources of bias, and this information used to inform future item writing.

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

Table 29 Mathematics 1 Items Flagged with C Category DIF

Category	Item	*b* Value	Group Preferred	MH DIF Value	*p* Value (significance < 0.001)
English as a First Language	733187	2.1344	Non-English	2.70	0.0000
	732935	-0.0661	Non-English	1.79	0.0000
	733127	1.4754	Non-English	1.53	0.0001
	733252	-1.9230	English	-1.56	0.0000

Biology

The DIF results are reported in Table 30 for Biology items that showed Category C DIF. These are items where there is a significant difference in the performance of candidates in different demographic groups.

Six items were flagged as showing DIF by first language, with three favouring candidates who did not speak English as a first language and three favouring candidates with English as their first language. These items are across a range of item difficulties. These items will be reviewed to identify likely sources of bias, and this information used to inform future item writing.

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

Table 30 Biology Items Flagged with C Category DIF

Category	Item	*b* Value	Group Preferred	MH DIF Value	*p* Value (significance < 0.001)
English as a First Language	731068	1.5825	Non-English	2.36	0.0000
	730984	-0.5006	Non-English	1.91	0.0000
	731274	-0.5111	Non-English	1.65	0.0000
	731029	-1.7905	English	-1.69	0.0001
	731031	1.2084	English	-1.88	0.0000
	730933	-0.4387	English	-1.88	0.0000

Chemistry

The DIF results are reported in Table 31 for Chemistry items that showed Category C DIF. These are items where there is a significant difference in the performance of candidates in different demographic groups.

Two items were flagged as showing DIF by first language, with both favouring candidates who speak English as a first language. These items are across a range of item difficulties, with one being very easy and one very difficult. The items will be reviewed to identify likely sources of bias, and this information used to inform future item writing.

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

Table 31 Chemistry Items Flagged with C Category DIF

Category	Item	*b* Value	Group Preferred	MH DIF Value	*p* Value (significance < 0.001)
English as a First Language	731806	1.4303	English	-1.95	0.0000
	731812	-2.0844	English	-2.71	0.0000

Physics

The DIF results are reported in Table 32 for Physics 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 candidates who identified as “Man” finding the item easier than those who identified as “Woman”. Three items were flagged as showing DIF by first language, with two items favouring candidates who did not speak English as a first language and one favouring candidates with English as their first language. These items are across a range of

item difficulties. The items will be reviewed to identify likely sources of bias, and this information used to inform future item writing.

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

Table 32 Physics Items Flagged with C Category DIF

Category	**Item**	***b* Value**	**Group Preferred**	**MH DIF Value**	***p* Value** (significance < 0.001)
Gender	747577	0.4911	Man	-1.77	0.0000
English as a First Language	732796	0.7803	Non- English	1.91	0.0000
	732799	-0.2413	Non- English	1.66	0.0000
	732735	-1.2080	English	-2.19	0.0000

Mathematics 2

There were no items on Mathematics 2 that showed significant Category C DIF.

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.