University Admissions Tests UK

UAT-UK Ltd

Annual Reports 2025-2026

ESAT 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 Engineering and Science Admissions Test (ESAT) was administered in two windows: 9th and 10th October 2025, and 6th and 7th January 2026, for candidates applying to start university in 2026. The test is modular, with all candidates taking Mathematics 1 and then a selection of the remaining modules (Biology, Chemistry, Physics and Mathematics 2). This report covers the 14,193 ESAT candidates (10,116 in October 2025 and 4,077 in January 2026) 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. This difference was 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 for the forms. The overall difficulty for Mathematics 2 and Physics were decreased following a review of the ESAT in the previous administration. This led to these modules being better targeted at the candidate ability than they were in the previous cycle.

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 but this was reduced compared to the previous cycle due to slight change in difficulty.

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.

In conclusion, the results of the 2025/26 ESAT administration were good. The test demonstrated good reliability, low measurement error, and balanced difficulties across the forms.

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.

The ESAT is available in two sittings per admissions cycle. This report covers those candidates who took the test in October 2025 (9th and 10th) and January 2025 (6th and 7th). 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 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 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 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 access arrangements 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 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 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 rank the 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 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 the 2024/25 cycle. 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 events.

Table 2 Ability Estimates Used to Scale the ESAT Modules

Event	Module	Percentile	Ability	Scaled Score
Oct 2025	Maths 1	50	0.4049	4.5
		90	2.7701	7.0
	Biology	50	0.5970	4.5
		90	1.6789	7.0
	Chemistry	50	0.2960	4.5
		90	1.9269	7.0
	Physics	50	0.3454	4.5
		90	1.8273	7.0
	Maths 2	50	0.0159	4.5
		90	1.6918	7.0

Table 3 Scaling Constants

Module	Constant	Multiplier
Maths 1	4.0720	1.0570
Biology	3.1205	2.3107
Chemistry	4.0463	1.5329
Physics	3.9173	1.6870
Maths 2	4.4763	1.4917

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. 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 74% of the candidates.

Table 4 ESAT Module Combinations

Combination	*N*	%
Maths 1, Maths 2, Physics	10,530	74%
Maths 1, Biology, Chemistry	1,334	9%
Maths 1, Maths 2, Chemistry	1,155	8%
Maths 1, Chemistry, Physics	642	5%
Maths 1, Maths 2	303	2%
Maths 1, Biology, Maths 2	110	1%
Maths1, Biology, Physics	103	1%
Other	16	0%

There were 14,193 ESAT candidates in total (all candidates take Maths 1), with 10,116 (71%) sitting in October 2025 and 4,077 (29%) in the January 2026 event (Table 5). This is an increase from the 2024/25 cycle, where there were 11,919 ESAT Maths 1 candidates, with 77% sitting in October 2024 and 23% in January 2025. This change is due to additional courses using the tests, an increase in candidate applications to the existing courses, and an increased proportion of applicants registering for the tests. 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. 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 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	14,193	4.58	1.47	1.0	9.0	3.6	4.4	5.3	6.7
	Oct	10,116	4.73	1.59	1.0	9.0	3.6	4.5	5.6	7.0
	Jan	4,077	4.19	1.03	1.0	9.0	3.6	4.1	4.7	5.4
Biology	All	1,547	4.44	1.87	1.0	9.0	3.2	4.5	5.7	7.0
	Oct	1,531	4.46	1.86	1.0	9.0	3.2	4.5	5.7	7.0
	Jan	16	NA	NA	NA	NA	NA	NA	NA	NA
Chemistry	All	3,146	4.61	1.74	1.0	9.0	3.5	4.5	5.7	7.0
	Oct	2,814	4.72	1.72	1.0	9.0	3.5	4.5	5.7	7.0
	Jan	332	3.63	1.62	1.0	8.9	2.5	3.5	4.7	5.7
Physics	All	11,290	4.25	1.71	1.0	9.0	3.0	4.1	5.3	6.5
	Oct	7,717	4.58	1.70	1.0	9.0	3.4	4.5	5.6	7.0
	Jan	3,573	3.55	1.51	1.0	9.0	2.5	3.4	4.6	5.5
Maths 2	All	12,113	4.40	1.63	1.0	9.0	3.3	4.2	5.4	6.5
	Oct	8,082	4.73	1.66	1.0	9.0	3.6	4.5	5.7	7.0
	Jan	4,031	3.75	1.33	1.0	9.0	2.9	3.7	4.6	5.4

Figure 1 Mathematics 1 Binned Scaled Score Distribution

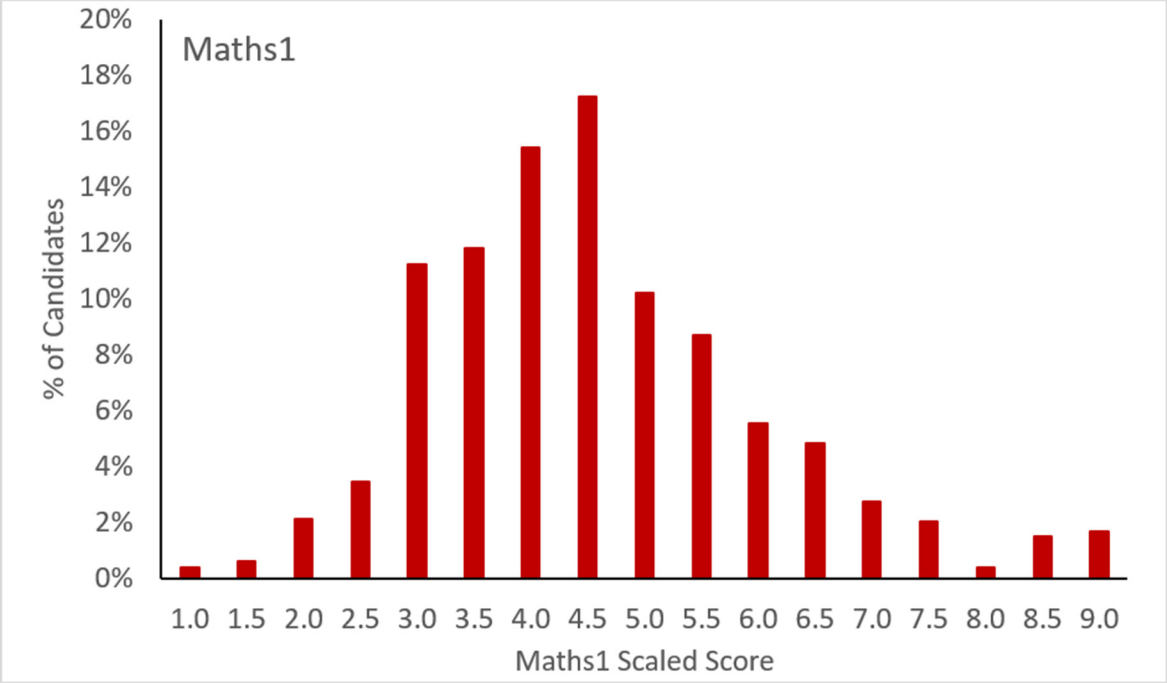


Figure 2 Biology Binned Scaled Score Distribution

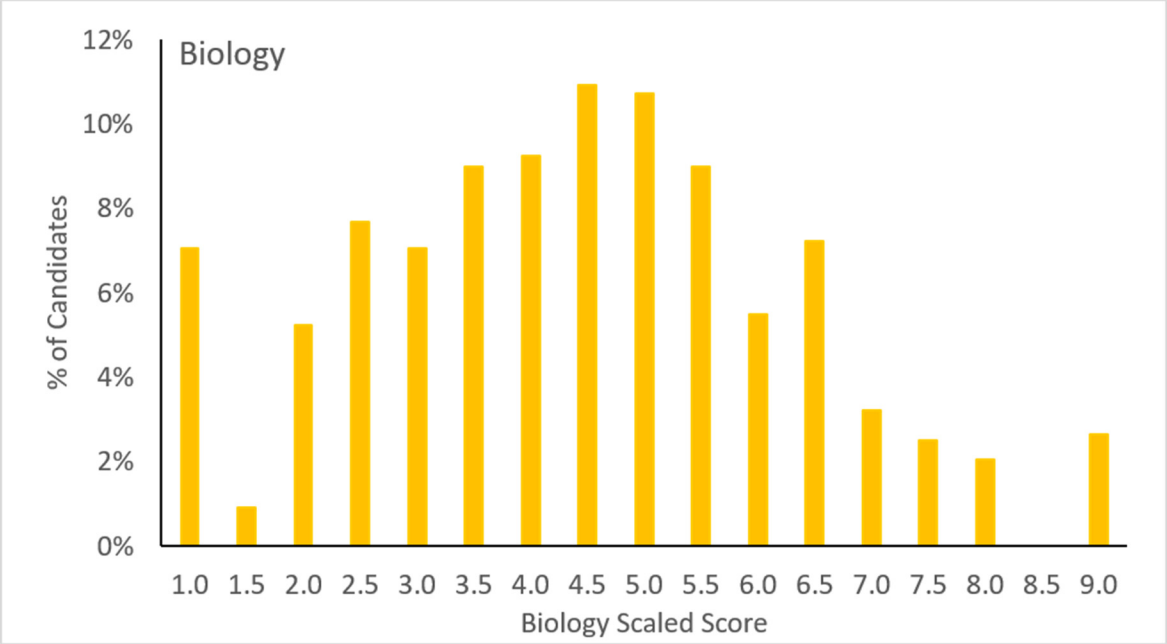


Figure 3 Chemistry Binned Scaled Score Distribution

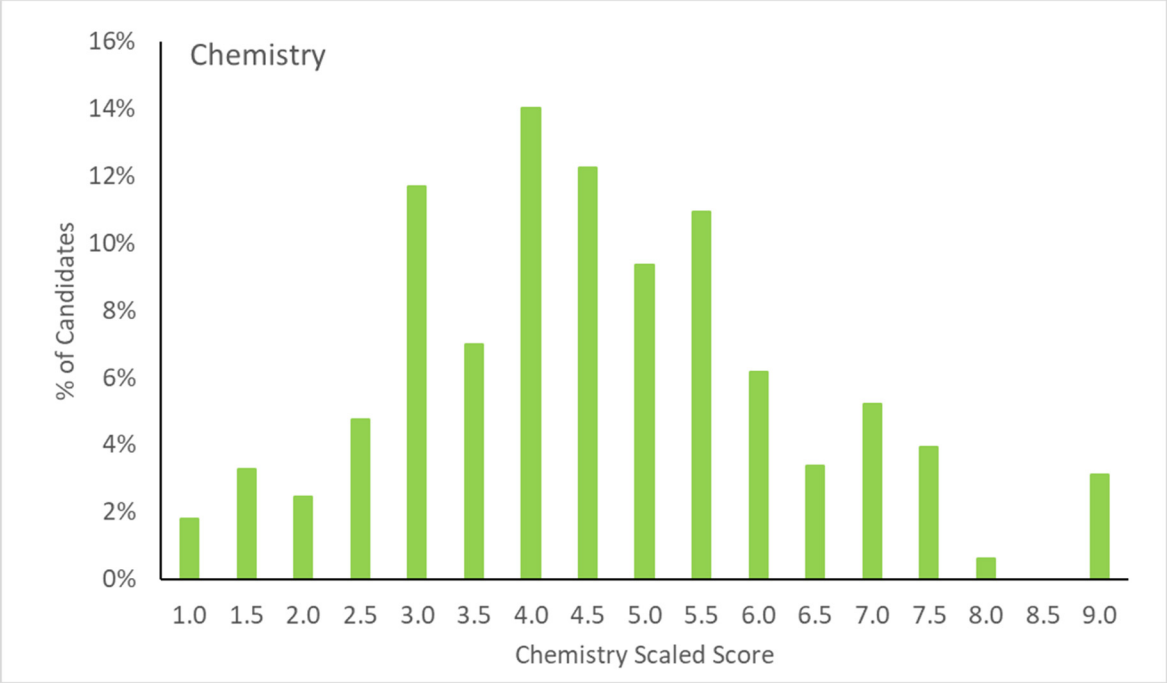


Figure 4 Physics Binned Scaled Score Distribution

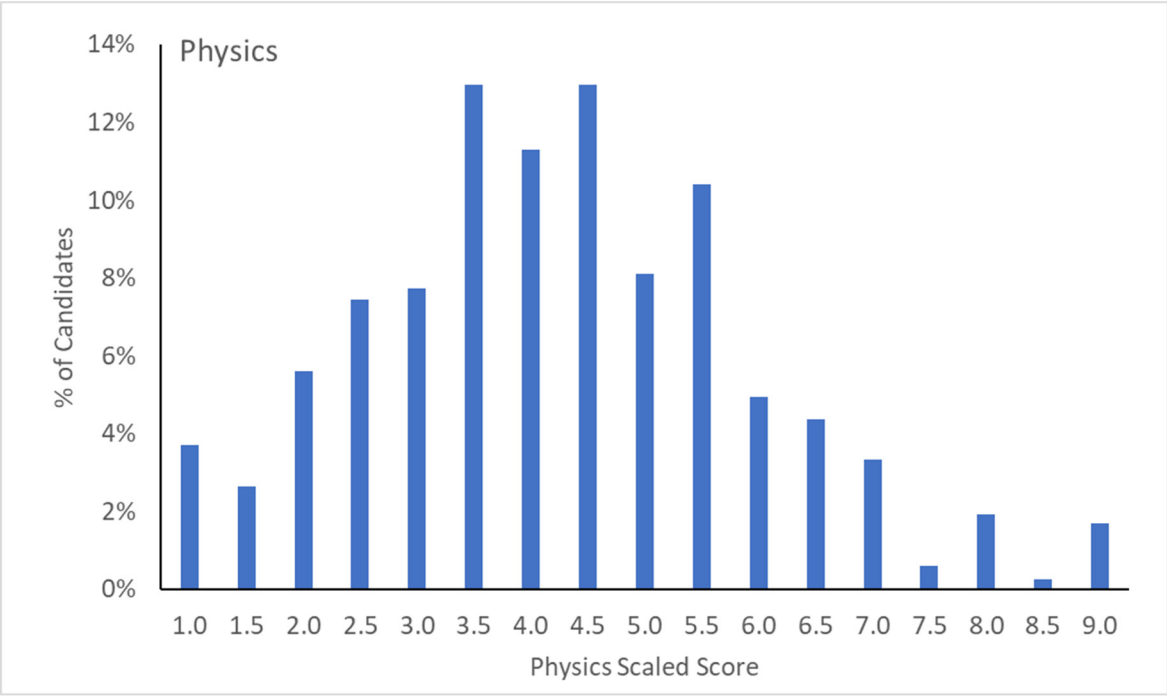
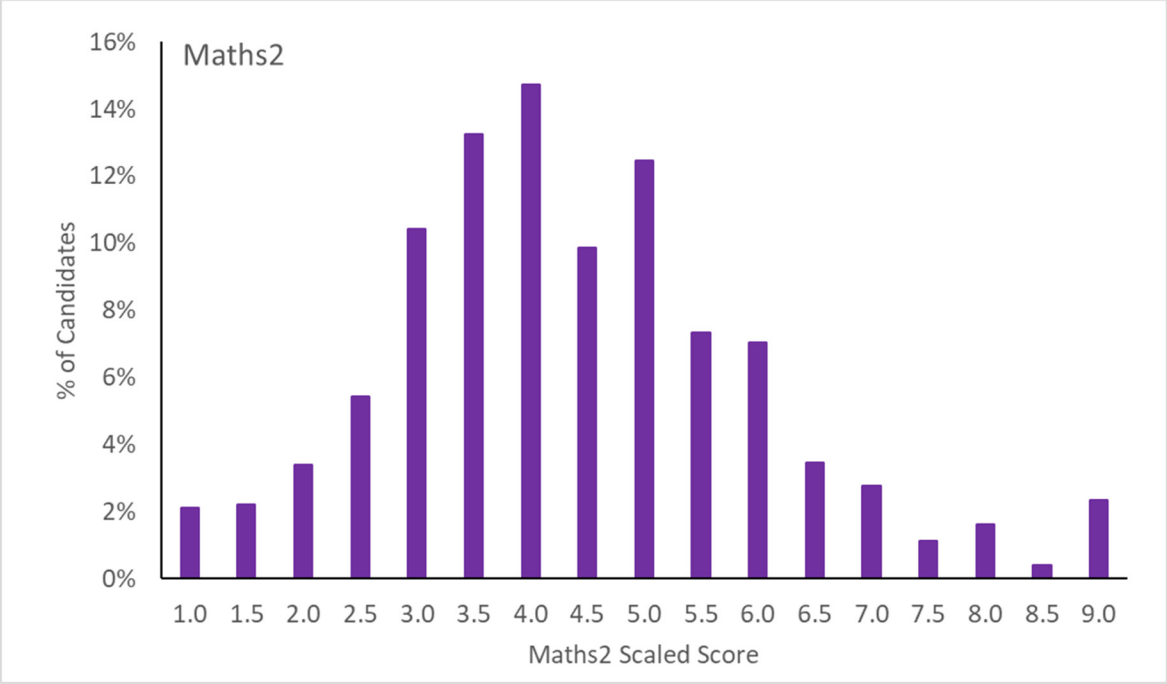


Figure 5 Mathematics 2 Binned Scaled Score Distribution



4.2 Test Results by Demographic Variables

4.2.1 Variation by Demographic Group

Pearson 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 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 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), 69% identified as "Man" and 30% as "Woman". Biology has the highest proportion of female candidates at 60% compared to only 24% for Physics. This is consistent with last year.

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 0.57 for Chemistry to 0.25 for Maths 2. 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

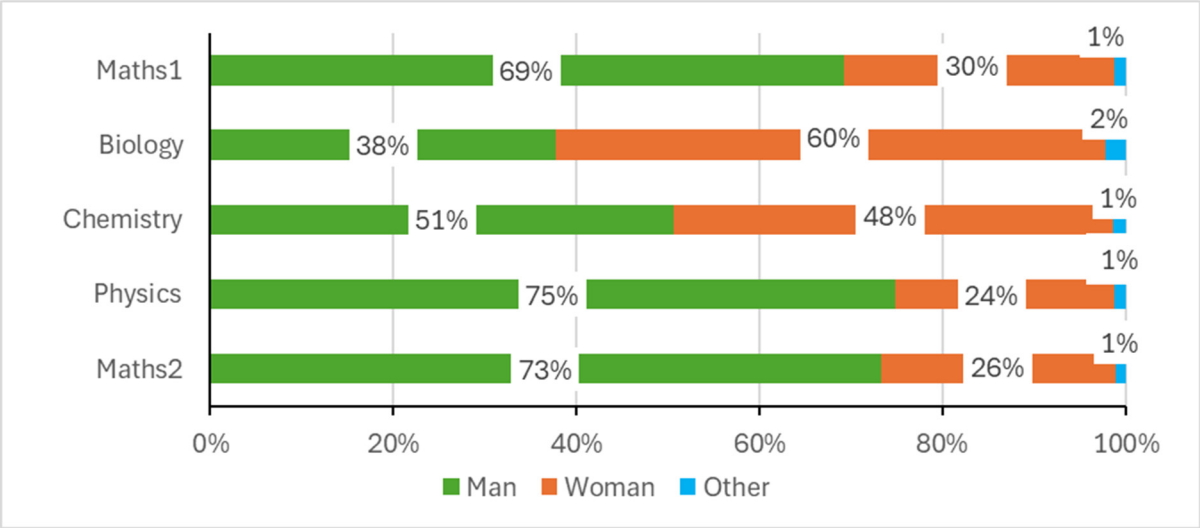
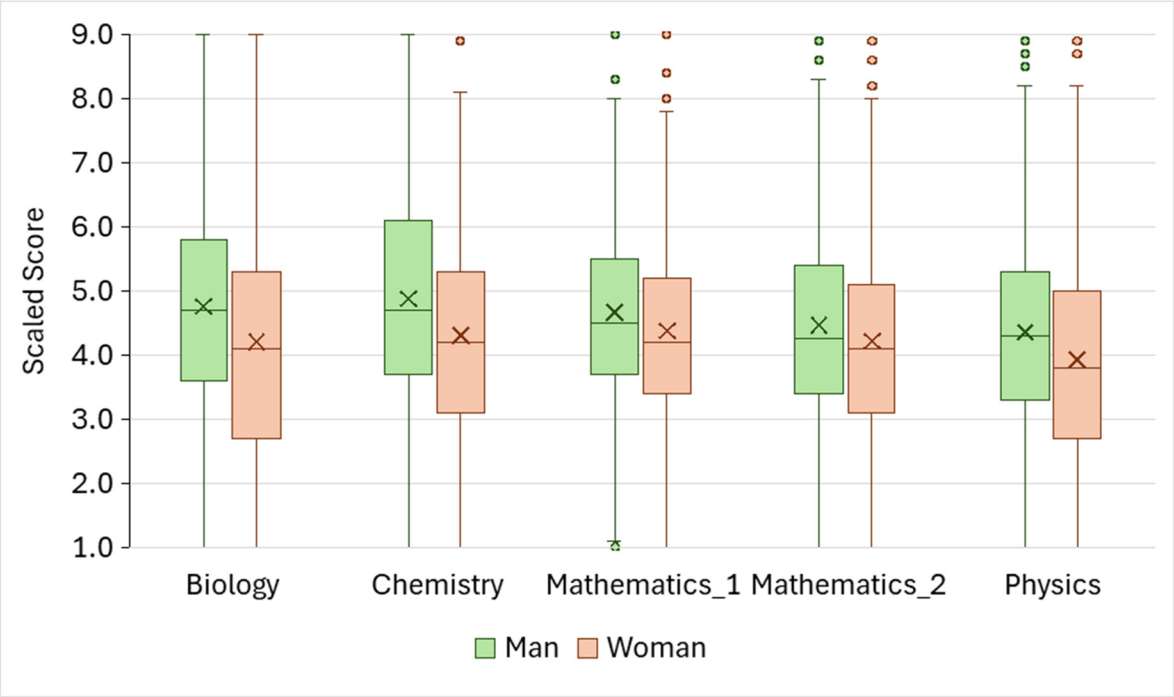


Table 7 Scaled Score Summary Statistics by Gender

Module	Gender	*N*	*%N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Man	9820	69%	4.66	1.46	1.0	9.0	3.7	4.5	5.5	6.7
	Woman	4195	30%	4.37	1.48	1.0	9.0	3.4	4.2	5.2	6.3
Biology	Man	585	38%	4.75	1.80	1.0	9.0	3.6	4.7	5.8	7.0
	Woman	929	60%	4.21	1.87	1.0	9.0	2.7	4.1	5.3	6.7
Chemistry	Man	1595	51%	4.88	1.77	1.0	9.0	3.7	4.7	6.1	7.0
	Woman	1507	48%	4.31	1.66	1.0	9.0	3.1	4.2	5.3	6.5
Physics	Man	8450	75%	4.36	1.70	1.0	9.0	3.3	4.3	5.3	6.5
	Woman	2705	24%	3.93	1.69	1.0	9.0	2.7	3.8	5.0	6.1
Maths 2	Man	8874	73%	4.47	1.62	1.0	9.0	3.4	4.3	5.4	6.6
	Woman	3100	26%	4.22	1.63	1.0	9.0	3.1	4.1	5.1	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 — 48% (Maths 2) to 51% (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 EU students are typically the weakest, with UK candidates in the middle and candidates outside of Europe being the strongest. For Biology, UK (mean of 4.43) and Other (mean of 4.54) candidates are broadly comparable, but for the other modules, Other candidates outperform UK candidates.

Figure 8 Percent of Candidates by Area of Residence

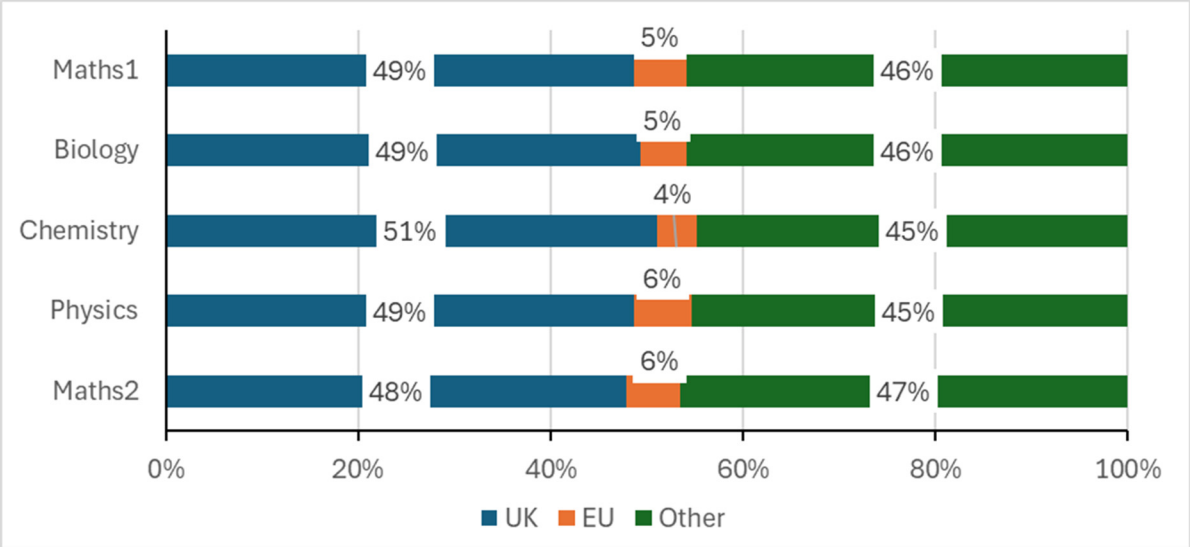
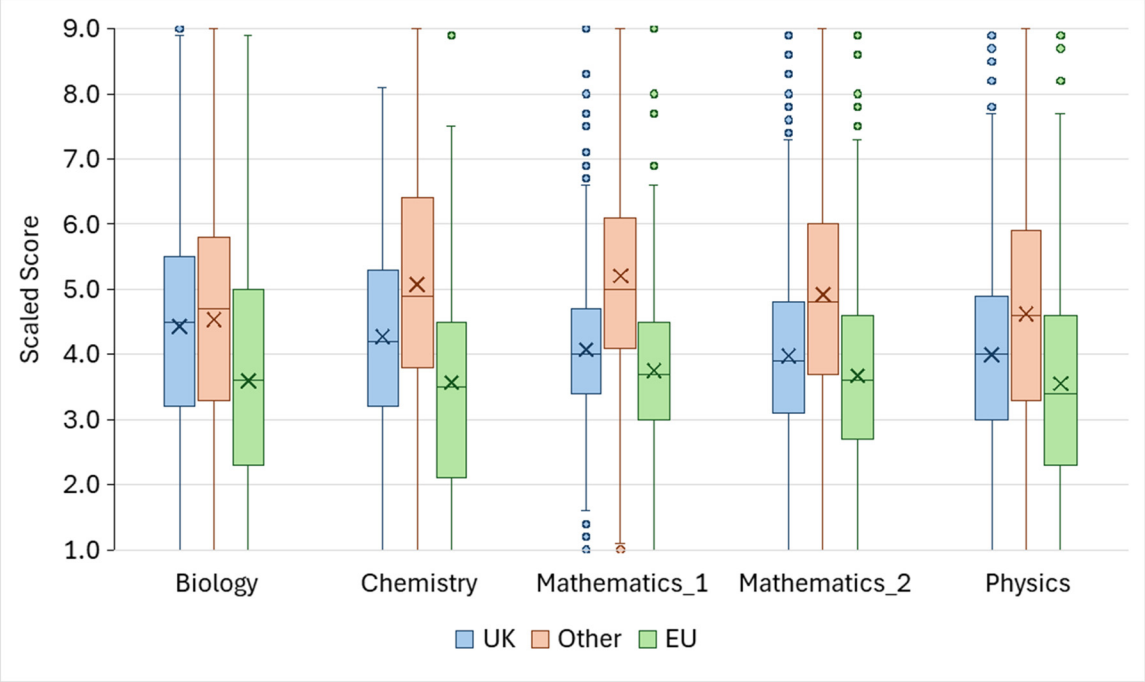


Table 8 Scaled Score Summary Statistics by Area of Residence

Module	Area	*N*	*%N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Maths 1	UK	6,909	49%	4.07	1.09	1.0	9.0	3.4	4.0	4.7	5.4
	EU	777	5%	3.75	1.27	1.0	9.0	3.0	3.7	4.5	5.3
	Other	6,507	46%	5.21	1.59	1.0	9.0	4.1	5.0	6.1	7.6
Biology	UK	763	49%	4.43	1.81	1.0	9.0	3.2	4.5	5.5	6.7
	EU	75	5%	3.60	1.80	1.0	8.9	2.3	3.6	5.0	6.0
	Other	709	46%	4.54	1.92	1.0	9.0	3.4	4.7	5.8	7.0
Chemistry	UK	1,605	51%	4.28	1.50	1.0	9.0	3.2	4.2	5.3	6.2
	EU	134	4%	3.57	1.75	1.0	8.9	2.1	3.5	4.5	6.1
	Other	1,407	45%	5.08	1.87	1.0	9.0	3.8	4.9	6.4	7.7
Physics	UK	5,504	49%	4.00	1.48	1.0	9.0	3.0	4.0	4.9	5.9
	EU	662	6%	3.56	1.66	1.0	8.9	2.3	3.4	4.6	5.7
	Other	5,124	45%	4.62	1.86	1.0	9.0	3.3	4.6	5.9	7.1
Maths 2	UK	5,801	48%	3.98	1.32	1.0	9.0	3.1	3.9	4.8	5.7
	EU	677	6%	3.68	1.51	1.0	9.0	2.7	3.6	4.6	5.4
	Other	5,635	47%	4.92	1.76	1.0	9.0	3.7	4.8	6.0	7.2

Figure 9 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 9.

Table 9 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 11 shows the breakdown of candidates by ethnicity for the ESAT modules. The largest ethnic group amongst UK candidates was White (42% for Maths 2 to 67% for Biology), closely followed by Asian (17% for Biology to 37% 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 (58%) of UK candidates were non-White.

Figure 10 Percent of UK Candidates by Ethnicity

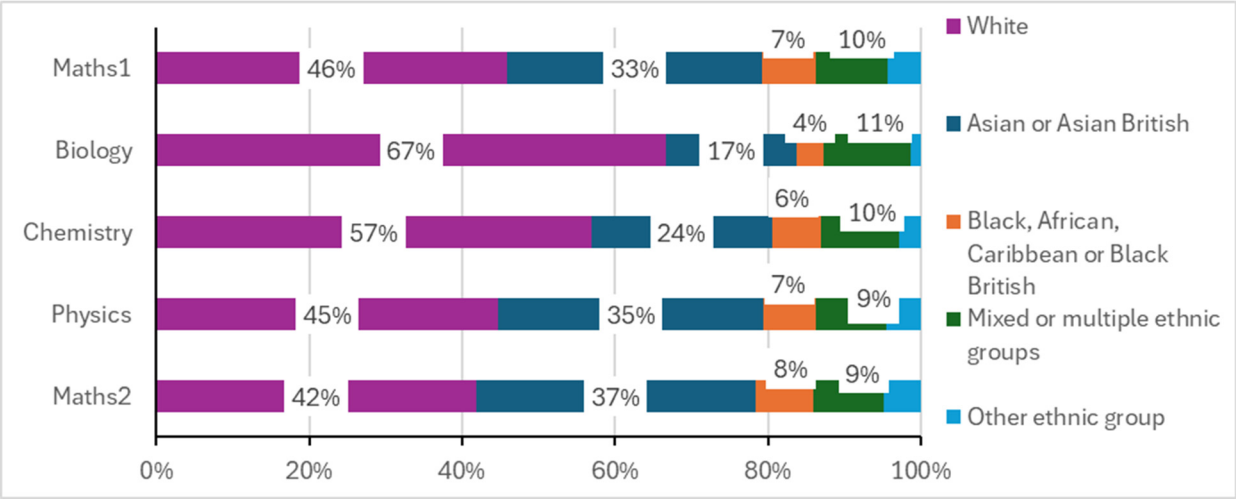


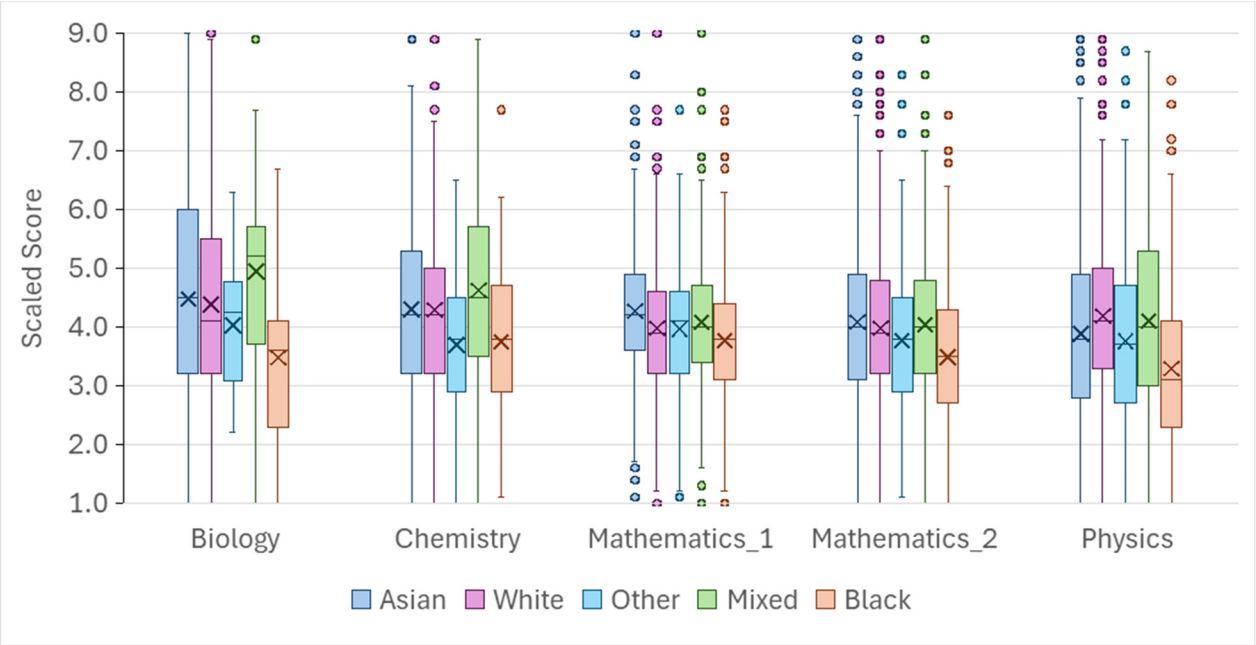
Table 10 summarises the scaled score summary statistics by ethnic group for UK students across the modules, for groups with at least 50 candidates. The differences in scaled score across the ethnic groups in the UK tend to reflect underlying trends within the UK.

Figure 11 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 4.27 and 4.08, respectively. For Biology and Chemistry, the strongest UK ethnic group was Mixed, but for Physics, the strongest group was White. The poorest performing UK ethnic group was those identifying as Black. This is in line with national trends.

Table 10 Scaled Score Summary Statistics by Ethnicity for UK Candidates

Module	Ethnicity	*N*	%*N*	Scaled Score				Percentile			
				Mean	S*D*	Min	Max	25	50	75	90
Maths 1	Asian	2,302	33%	4.27	1.16	1.1	9.0	3.6	4.2	4.9	5.7
	Black	479	7%	3.78	1.01	1.0	7.8	3.1	3.8	4.4	5.0
	Mixed	658	10%	4.08	1.12	1.0	9.0	3.4	4.0	4.7	5.5
	Other	297	4%	3.97	1.02	1.1	7.7	3.2	4.1	4.6	5.2
	White	3173	46%	3.98	1.04	1.0	9.0	3.2	3.9	4.6	5.3
Biology	Asian	130	17%	4.48	1.93	1.0	9.0	3.2	4.5	6.0	7.0
	Black	27	4%	NA	NA	NA	NA	NA	NA	NA	NA
	Mixed	87	11%	4.94	1.90	1.0	9.0	3.7	5.2	5.7	7.7
	Other	10	1%	NA	NA	NA	NA	NA	NA	NA	NA
	White	509	67%	4.39	1.77	1.0	9.0	3.2	4.1	5.5	6.7
Chemistry	Asian	378	24%	4.30	1.62	1.0	9.0	3.2	4.2	5.3	6.2
	Black	103	6%	3.75	1.24	1.1	7.7	2.9	3.8	4.7	5.3
	Mixed	163	10%	4.62	1.65	1.0	8.9	3.5	4.5	5.7	6.9
	Other	47	3%	NA	NA	NA	NA	NA	NA	NA	NA
	White	914	57%	4.30	1.44	1.0	8.9	3.2	4.2	5.0	6.2
Physics	Asian	1,911	35%	3.89	1.54	1.0	9.0	2.8	3.8	4.9	5.9
	Black	371	7%	3.29	1.40	1.0	8.2	2.3	3.1	4.1	5.0
	Mixed	514	9%	4.11	1.46	1.0	8.7	3.0	4.0	5.3	5.9
	Other	247	4%	3.76	1.51	1.0	8.7	2.7	3.7	4.7	5.6
	White	2461	45%	4.19	1.40	1.0	9.0	3.3	4.1	5.0	6.1
Maths 2	Asian	2,119	37%	4.08	1.43	1.0	9.0	3.1	4.0	4.9	5.8
	Black	441	8%	3.49	1.21	1.0	7.6	2.7	3.5	4.3	5.1
	Mixed	533	9%	4.04	1.30	1.0	8.9	3.2	4.0	4.8	5.8
	Other	284	5%	3.77	1.26	1.1	8.3	2.9	3.8	4.5	5.4
	White	2424	42%	3.98	1.24	1.0	9.0	3.2	3.9	4.8	5.4

Figure 11 Box and Whisker Plot of Scaled Score by Ethnicity



4.2.5 First Language

Figure 12 illustrates the proportion of candidates who identified English as their first language by module. The proportion of candidates with English as a first language by modules was broadly similar, ranging from 53% to 56%. This is in contrast to last year, when the proportion of candidates with English as a first language for Maths 2 was 65%. Interestingly, candidates with English as a first language were weaker than those whose first language was not English (Table 11) across all modules.

Figure 12 Percent of Candidates by First Language

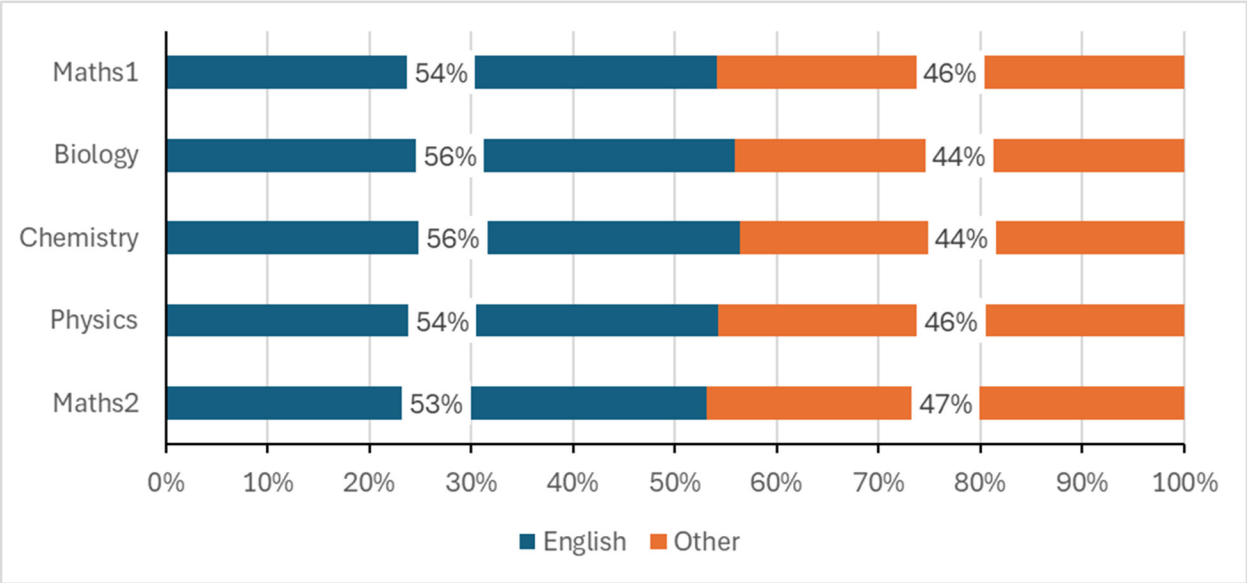
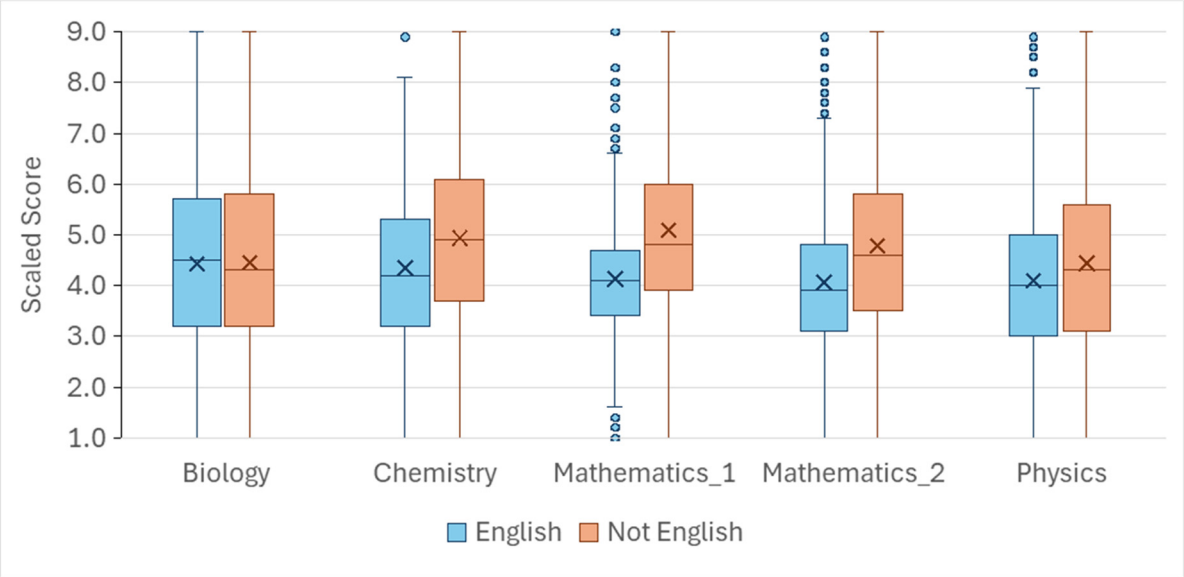


Table 11 Scaled Score Summary Statistics by First Language

Module	First Language	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Maths 1	English	7,676	54%	4.14	1.16	1.0	9.0	3.4	4.1	4.7	5.6
	Other	6,517	46%	5.09	1.62	1.0	9.0	3.9	4.8	6.0	7.6
Biology	English	8,64	56%	4.43	1.90	1.0	9.0	3.2	4.5	5.7	7.0
	Other	6,83	44%	4.45	1.83	1.0	9.0	3.2	4.3	5.8	7.0
Chemistry	English	1,774	56%	4.35	1.66	1.0	9.0	3.2	4.2	5.3	6.5
	Other	1,372	44%	4.93	1.79	1.0	9.0	3.7	4.9	6.1	7.5
Physics	English	6,131	56%	4.10	1.56	1.0	9.0	3.0	4.0	5.0	6.1
	Other	5,159	44%	4.44	1.85	1.0	9.0	3.1	4.3	5.6	7.1
Maths 2	English	6,438	53%	4.06	1.38	1.0	9.0	3.1	3.9	4.8	5.8
	Other	5,675	47%	4.78	1.79	1.0	9.0	3.5	4.6	5.8	7.2

Figure 13 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 0.02 higher than those with English as a first language, and largest for Maths 1 (0.96). This result is primarily due to a large number of candidates in China who are very strong at Maths and Physics. It is, however, especially interesting that even for those subjects with a higher reading load, such as Biology, candidates for whom English was not their first language performed better than those who did speak English as a first language.

Figure 13 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 37% (Biology) to 41% (Physics and Maths 2) of candidates, was a Further Education College or Sixth Form College (Figure 14). The least common school type in the UK, excluding University, was an Academy/Secondary School, with 17% to 19% of candidates selecting these school types across the modules. Grammar school candidates made up 17% to 21% 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 100 candidates can be found in Table 12. Figure 15 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 14 Percent of Candidates by School Type

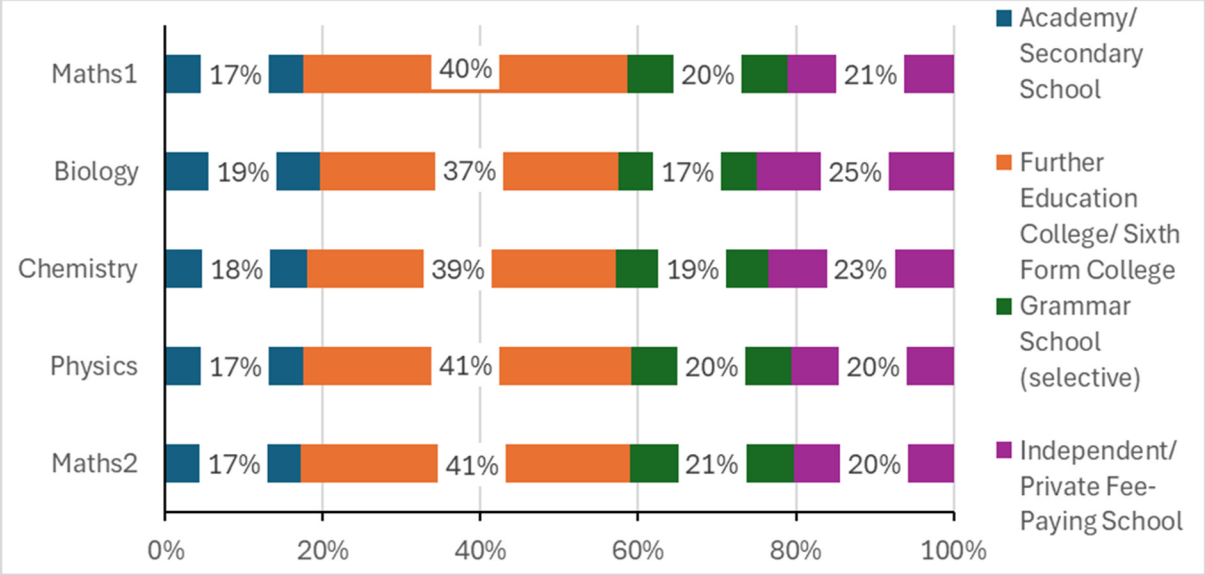
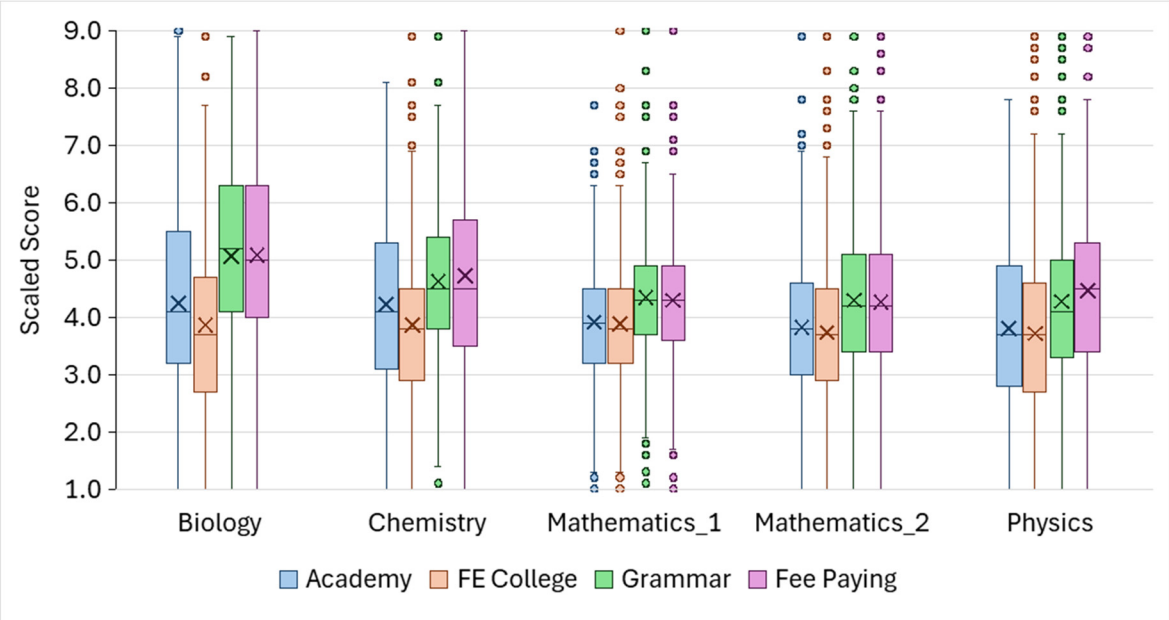


Table 12 Scaled Score Summary Statistics by School Type

Module	School Type	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Secondary School	1,192	17%	3.91	1.06	1.0	7.8	3.2	3.9	4.5	5.2
	FE/ 6th Form College	2,795	40%	3.89	1.00	1.0	9.0	3.2	3.8	4.5	5.1
	Grammar School	1,381	20%	4.34	1.11	1.1	9.0	3.7	4.3	4.9	5.7
	Private Fee-Paying School	1,433	21%	4.29	1.16	1.0	9.0	3.6	4.3	4.9	5.7
Biology	Secondary School	147	19%	4.25	1.76	1.0	9.0	3.2	4.1	5.5	6.7
	FE/ 6th Form College	285	37%	3.87	1.63	1.0	8.9	2.7	3.7	4.7	6.0
	Grammar School	131	17%	5.07	1.60	1.0	8.9	4.1	5.2	6.3	7.0
	Private Fee-Paying School	188	25%	5.09	1.86	1.0	9.0	4.0	5.0	6.3	7.7
Chemistry	Secondary School	285	18%	4.23	1.51	1.0	8.9	3.1	4.1	5.3	6.1
	FE/ 6th Form College	623	39%	3.86	1.35	1.0	8.9	2.9	3.8	4.5	5.7
	Grammar School	306	19%	4.63	1.41	1.1	8.9	3.8	4.5	5.4	6.5
	Private Fee-Paying School	374	23%	4.73	1.57	1.0	9.0	3.5	4.5	5.7	6.9
Physics	Secondary School	951	17%	3.80	1.45	1.0	9.0	2.8	3.7	4.9	5.7
	FE/ 6th Form College	2,251	41%	3.71	1.41	1.0	8.9	2.7	3.7	4.6	5.3
	Grammar School	1,101	20%	4.28	1.46	1.0	8.9	3.3	4.1	5.0	6.2
	Private Fee-Paying School	1,111	20%	4.47	1.46	1.0	9.0	3.4	4.5	5.3	6.2
Maths 2	Secondary School	982	17%	3.83	1.28	1.0	9.0	3.0	3.8	4.6	5.4
	FE/ 6th Form College	2,383	41%	3.73	1.25	1.0	9.0	2.9	3.7	4.5	5.1
	Grammar School	1,194	21%	4.30	1.33	1.0	9.0	3.4	4.2	5.1	6.0
	Private Fee-Paying School	1,152	20%	4.27	1.36	1.0	8.9	3.4	4.2	5.1	6.0

Figure 15 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. Between 7% (Biology) and 10% (Maths 1, Maths 2 and Physics) of the candidates who took ESAT received free school meals (Figure 16). 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 accounts for the lower percentage of candidates selecting "No" than last year, but overall the proportion of candidates selecting "Yes" is consistent with last year. 7% to 10% receiving free school meals 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 17. 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 16 Percent of Candidates by Free School Meals (UK Only)

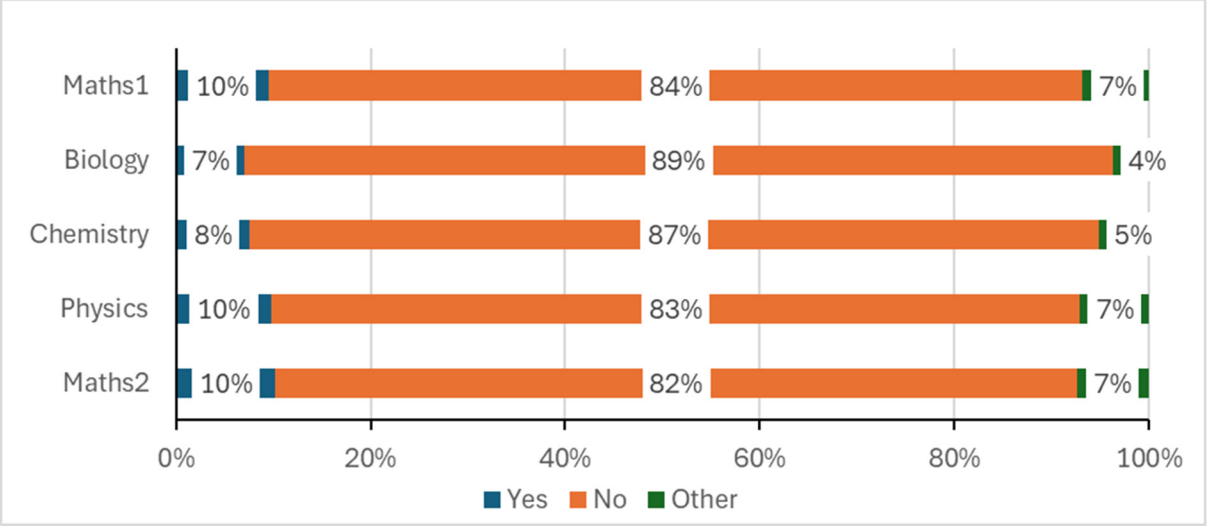
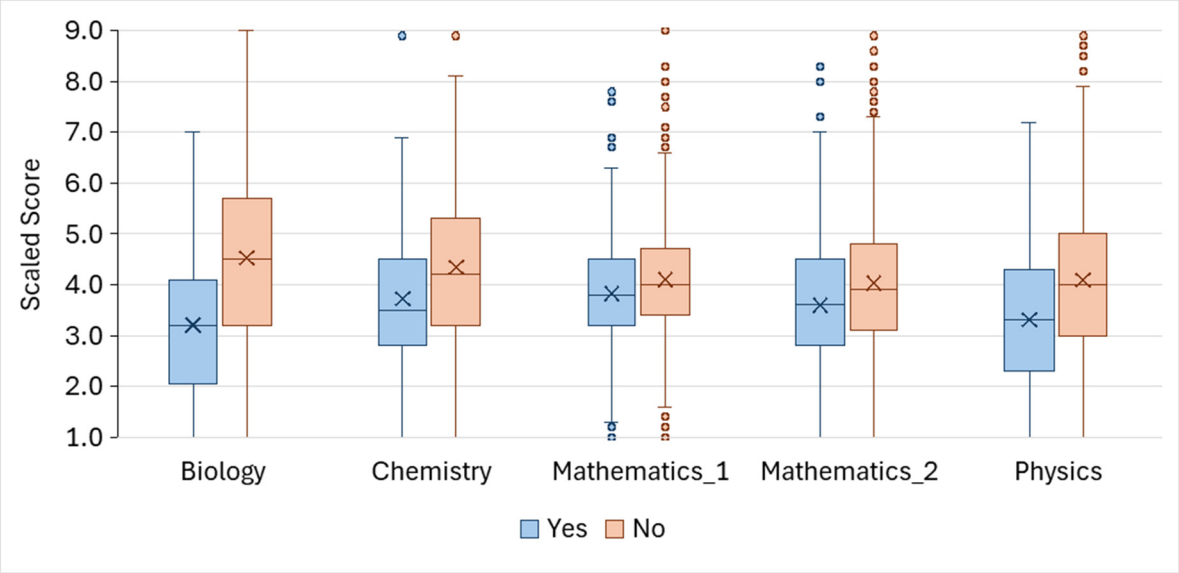


Table 13 Summary Statistics by Free School Meals (UK Only)

Module	Free School Meals	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Yes	659	10%	3.83	1.01	1.0	7.8	3.2	3.8	4.5	5.0
	No	5,781	84%	4.10	1.10	1.0	9.0	3.4	4.0	4.7	5.5
Biology	Yes	54	7%	3.21	1.51	1.0	7.0	2.1	3.2	4.1	5.2
	No	681	89%	4.53	1.79	1.0	9.0	3.2	4.5	5.7	7.0
Chemistry	Yes	122	8%	3.72	1.37	1.0	8.9	2.8	3.5	4.5	5.4
	No	1,400	87%	4.34	1.50	1.0	9.0	3.2	4.2	5.3	6.2
Physics	Yes	540	10%	3.31	1.34	1.0	7.2	2.3	3.3	4.3	5.0
	No	4,575	83%	4.09	1.47	1.0	9.0	3.0	4.0	5.0	6.1
Maths 2	Yes	591	10%	3.59	1.26	1.0	8.3	2.8	3.6	4.5	5.2
	No	4,785	82%	4.04	1.32	1.0	9.0	3.1	3.9	4.8	5.7

Figure 17 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 2 and Physics) and 83% (Biology) of candidates responded "Yes" (Figure 18). 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 19. These data show that across all modules, the candidates who had a parent/guardian attend higher education significantly outperformed those who did not.

Figure 18 Percent of Candidates by Parent Higher Education

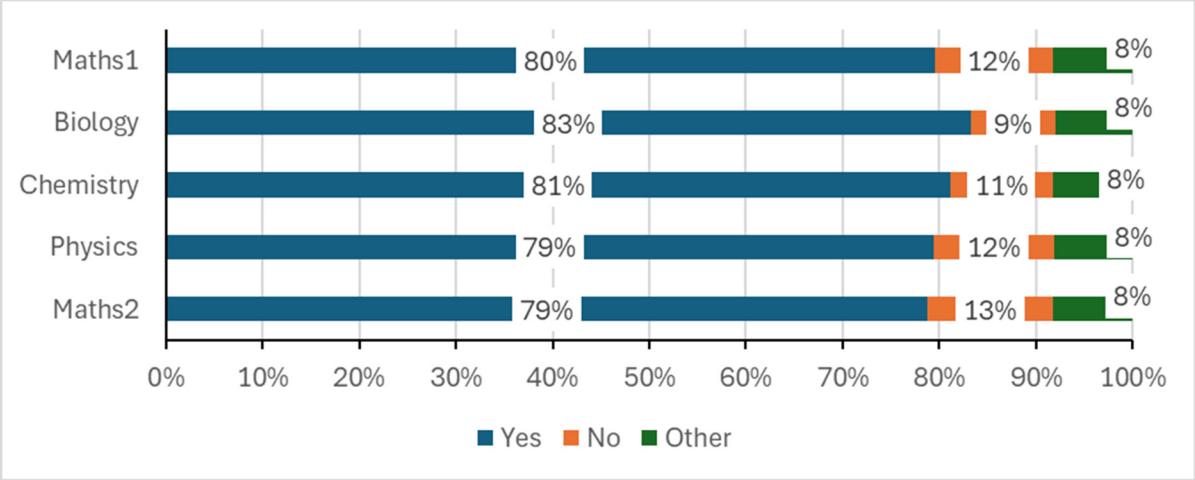
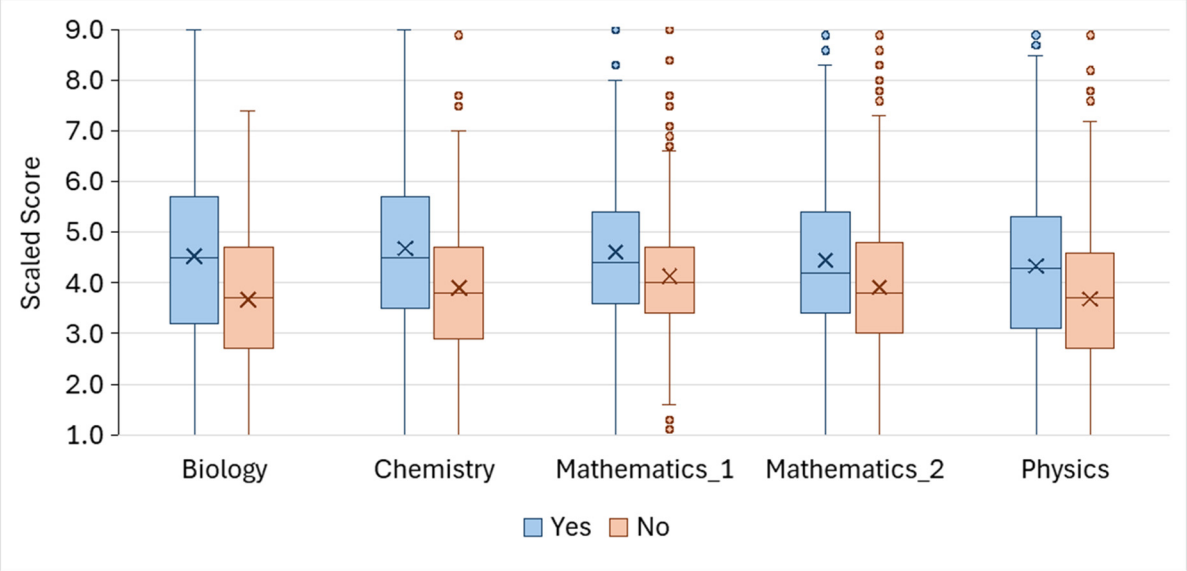


Table 14 Summary Statistics by Parent/Guardian Education

Module	Parent Higher Education	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Yes	11,287	80%	4.61	1.48	1.0	9.0	3.6	4.4	5.4	6.7
	No	1,739	12%	4.14	1.27	1.1	9.0	3.4	4.0	4.7	5.7
Biology	Yes	1,287	83%	4.53	1.89	1.0	9.0	3.2	4.5	5.7	7.0
	No	136	9%	3.67	1.55	1.0	7.4	2.7	3.7	4.7	5.7
Chemistry	Yes	2,552	81%	4.68	1.76	1.0	9.0	3.5	4.5	5.7	7.0
	No	333	11%	3.91	1.51	1.0	8.9	2.9	3.8	4.7	6.0
Physics	Yes	8,963	79%	4.34	1.71	1.0	9.0	3.1	4.3	5.3	6.5
	No	1,411	12%	3.69	1.53	1.0	8.9	2.7	3.7	4.6	5.6
Maths 2	Yes	9,543	79%	4.46	1.63	1.0	9.0	3.4	4.2	5.4	6.7
	No	1,566	13%	3.91	1.45	1.0	9.0	3.0	3.8	4.8	5.7

Figure 19 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, 85% to 89% of candidates said that they did not have such a health condition or disability (Figure 20). This is a slight drop from last year, as the candidates also had the option of "I prefer not to say" and "Don't know", which were not available last year. There is a difference in performance across the two groups, with those who responded "No" outperforming those who responded "Yes" (Table 15; Figure 21).

Figure 20 Percent of Candidates by Learning Difficulty/Chronic Health Condition

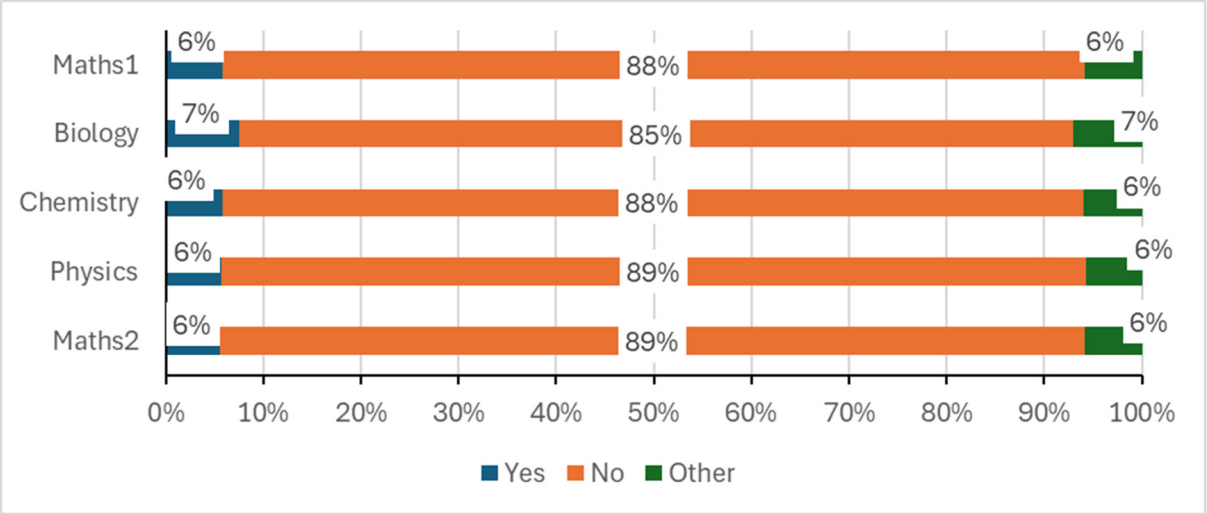
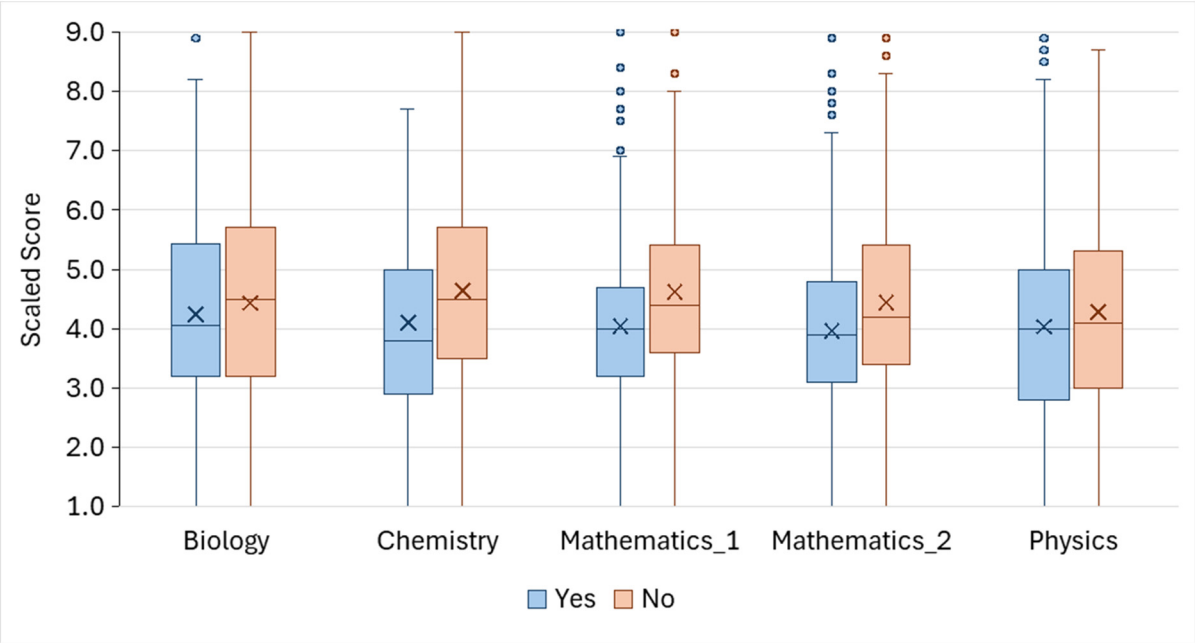


Table 15 Summary Statistics by Learning Difficulty/Chronic Health Condition

Module	Learning Difficulty	*N*	%*N*	Scaled Score				Percentile			
				Mean	*SD*	Min	Max	25	50	75	90
Maths 1	Yes	827	6%	4.04	1.18	1.0	9.0	3.2	4.0	4.7	5.4
	No	12,528	88%	4.62	1.49	1.0	9.0	3.6	4.4	5.4	6.7
Biology	Yes	116	7%	4.24	1.90	1.0	8.9	3.2	4.1	5.4	7.0
	No	1,322	85%	4.44	1.87	1.0	9.0	3.2	4.5	5.7	7.0
Chemistry	Yes	183	6%	4.10	1.65	1.0	8.9	2.9	3.8	5.0	6.2
	No	2,776	88%	4.64	1.75	1.0	9.0	3.5	4.5	5.7	7.0
Physics	Yes	646	6%	4.03	1.63	1.0	9.0	2.8	4.0	5.0	6.1
	No	10,004	89%	4.28	1.71	1.0	9.0	3.0	4.1	5.3	6.5
Maths 2	Yes	678	6%	3.96	1.40	1.0	8.9	3.1	3.9	4.8	5.7
	No	10,730	89%	4.44	1.64	1.0	9.0	3.4	4.2	5.4	6.7

Figure 21 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, 614 candidates had access arrangements for Maths 1, the compulsory module, which is around 5% of the candidate population. Candidates can apply for more than one access arrangement. Table 16 summarises the number of access arrangements by the type required, but note that as candidates can apply for more than one access arrangement, some are counted more than once. The most common request was for extra time and/or pause-the-clock, with 607 candidates requesting this.

Table 16 Number of Access Arrangements by Type

Access Arrangement	Maths 1	Biology	Chemistry	Physics	Maths 2
Extra Time	407	61	84	320	331
Pause-the-clock	60	12	18	41	47
Extra time + pause-the-clock	140	20	31	105	116
Separate Room	56	5	13	45	46
Other	70	11	20	52	54

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 Maths 1 were excellent, ranging from 0.68 to 0.84, with only one form below 0.70. Maths 1 performed the strongest in terms of internal consistency. For Biology, there was a much smaller candidate sample, and the reliabilities ranged from 0.65 to 0.69. This is slightly lower but also reflects a more homogeneous candidate pool with a lower raw score standard deviation. For Chemistry, reliabilities for the October forms were excellent and ranged from 0.76 to 0.80. For the January forms, there was a much smaller number of candidates, with fewer than 100 candidates, and therefore these values may not be representative. For Physics, the reliabilities ranged from 0.70 to 0.77, which is an improvement on last year, which can partly be attributed to the test being better targeted in terms of difficulty. This is more marked in Maths 2, where the reliabilities ranged from 0.66 to 0.81. This is an improvement on last year, where the reliabilities ranged from 0.53 to 0.77.

Table 17 Raw Score Reliability and SEM

Module	Event	Raw Score Reliability	
		Cronbach's Alpha	*SEM*
Maths 1	Oct 2025	0.81 to 0.84	1.89 to 2.21
	Jan 2026	0.68 to 0.79	2.15 to 2.21
Biology	Oct 2025	0.65 to 0.69	2.12 to 2.19
	Jan 2026	-	-
Chemistry	Oct 2025	0.76 to 0.80	2.02 to 2.12
	Jan 2026	-	-
Physics	Oct 2025	0.71 to 0.79	2.03 to 2.21
	Jan 2026	0.70 to 0.77	2.23 to 2.26
Maths 2	Oct 2025	0.75 to 0.81	2.02 to 2.22
	Jan 2026	0.66 to 0.71	2.16 to 2.21

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 2025	0.74 to 0.82	0.54 to 0.73
	Jan 2026	0.69 to 0.79	0.51 to 0.54
Biology	Oct 2025	0.66 to 0.70	1.03 to 1.05
	Jan 2026	-	-
Chemistry	Oct 2025	0.71 to 0.80	0.75 to 0.87
	Jan 2026	-	-
Physics	Oct 2025	0.73 to 0.76	0.77 to 0.95
	Jan 2026	0.71 to 0.78	0.73 to 0.76
Maths 2	Oct 2025	0.75 to 0.79	0.70 to 0.88
	Jan 2026	0.67 to 0.73	0.68 to 0.69

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 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 20). The mean 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 Maths 1 and Maths 2, and only fractionally below for the remining modules. The distribution of candidates by module time is illustrated in Figure 22, which shows that over 90% of candidates used between 35 and 40 minutes, with the exception of Biology, where 86% used between 35 and 40 minutes.

Table 19. Test Time Summary Statistics

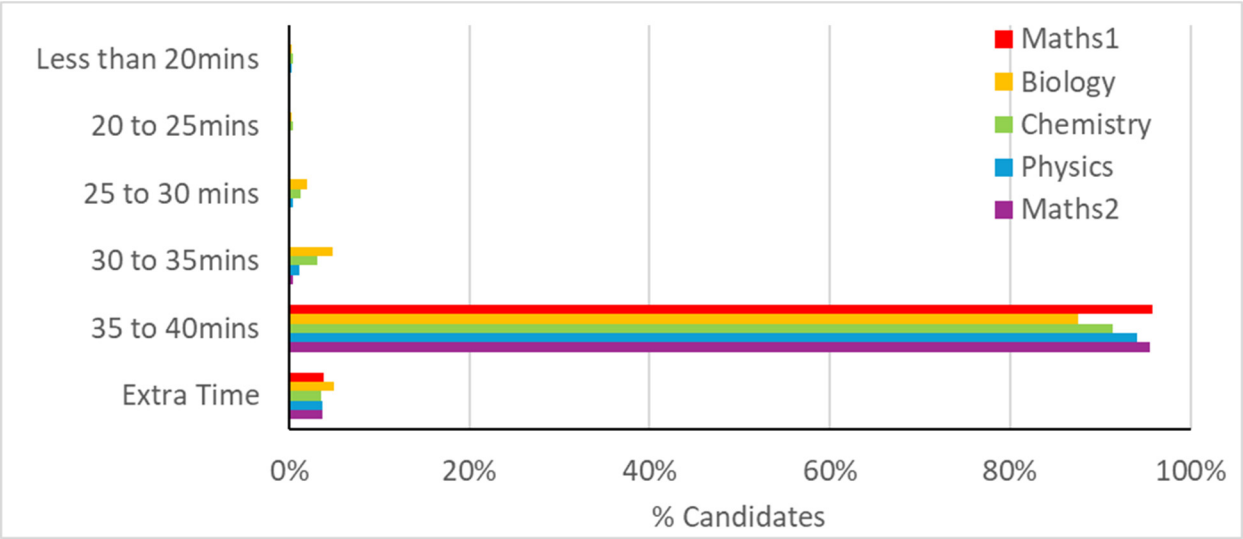
Module	*N*	Test Time				
		Mean	Median	*SD*	Min	Max
Maths 1	13,645	00:39:50	00:40:00	00:00:57	00:01:42	00:40:02
Biology	1,470	00:38:54	00:39:53	00:02:48	00:12:09	00:40:02
Chemistry	3,035	00:39:04	00:39:53	00:02:47	00:01:41	00:40:02
Physics	10,876	00:39:29	00:39:59	00:02:19	00:00:00	00:40:02
Maths 2	11,670	00:39:43	00:40:00	00:01:36	00:00:06	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	548	4%
Biology	77	5%
Chemistry	111	4%
Physics	414	4%
Maths 2	443	4%

Figure 22 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 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 (2% of candidates) and Chemistry (2% 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. In the 2024/25 cycle, 10% of candidates were not administered all items for Physics, but since the module has been made slightly easier (and better targeted at candidates), this has reduced to 7%, which would indicate that the test is borderline speeded. This could be reduced, if desired, by making the module slightly easier again. In the 2024/25 cycle, Maths 1 and Maths 2 were the most speeded, where 15% and 17% of candidates, respectively, had at least one unreached item. In the current cycle, this has reduced to 13% for Maths 1 and 14% for Maths 2. The reduction in Maths 2 can be partly attributed to making the test slightly easier and better targeted to the population. The speededness will continue to be monitored, but the reduction in the number of candidates having unreached items is a positive sign.

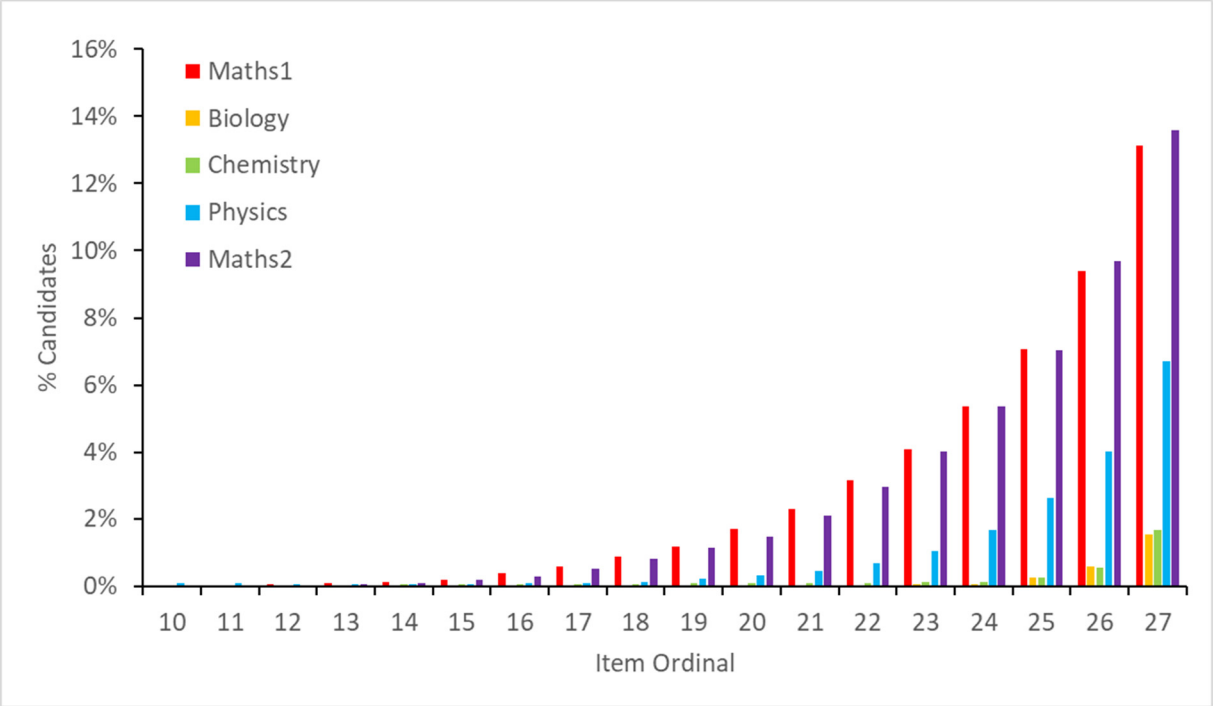
To reduce speededness, item time could be considered when building the forms in the future as long as the supply of items is large enough to support this. Item time statistics should also be reviewed to identify any trends amongst items that are taking more time.

Table 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	14,193	1892	13%	3.74	1 to 26
Biology	1,547	24	2%	1.63	1 to 5
Chemistry	3,146	52	2%	2.29	1 to 26
Physics	11,290	772	7%	2.79	1 to 27
Maths 2	12,113	1667	14%	3.60	1 to 25

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 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 23 shows that Maths 1 and Maths 2 are the most speeded and Biology and Chemistry are the least.

Figure 23 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 22 summarises the number of candidates with unreached or low time items. The trends observed very much mirror those observed in Table 21, with the percentage of candidates with low time/unreached items lowest for Biology (3%) and Chemistry (4%). The most speeded module is

still Maths 2, with 40% of candidates having at least one item unreached or below 5 seconds. This compares to over 50% in the 2024/25 cycle and is a significant reduction and shows that this module is functioning better since the difficulty was reduced. There has been a reduction in the percentage of candidates having unreached or low-time items across all the modules compared to last year, with a reduction in Maths 1 from 46% to 37% and a reduction from 31% to 20% in Physics. This will continue to be monitored.

Table 22 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	14,193	5,257	37%	3.44	1 to 26
Biology	1,547	46	3%	1.76	1 to 7
Chemistry	3,146	135	4%	2.35	1 to 26
Physics	11,290	2,288	20%	2.60	1 to 27
Maths 2	12,113	4,845	40%	3.41	1 to 26

6. Item Performance

Each year, Pearson 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 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 ESAT, 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 ESAT 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 universities 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 Mathematics 1 Item Performance

Of the items used in the Mathematics 1 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 0.12 to 0.59, with a mean of 0.38, which is excellent and comparable to the previous year.

Table 23 Mathematics 1 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	5%
	p Value less than 0.20	4%
	None	91%
Total		100%

6.1.2 Biology Item Performance

Of the items used in the Biology 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 0.09 to 0.49, with a mean of 0.33, which is excellent.

Table 24 Biology 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%

6.1.3 Chemistry Item Performance

Of the items used in the Chemistry forms, 100% of items met the criteria — well above the 80% target for new items. All items had a positive point biserial ranging from 0.24 to 0.56, with a mean of 0.39, which is excellent.

Table 25 Chemistry 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	9%
	p Value less than 0.20	4%
	None	87%
Total		100%

6.1.4 Physics Item Performance

Of the items used in the Physics forms, 98% of items met the criteria – well above the 80% target for new items. The point biserials ranged from -0.01 to 0.54 with a mean of 0.36, which is excellent.

Table 26 Physics 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	1%
Pass	*p* Value greater than 0.90	3%
	p Value less than 0.20	7%
	None	88%
Total		100%

6.1.5 Mathematics 2 Item Performance

Of the items used in the Mathematics 2 forms, 99% of the items met the criteria – well above the 80% typical target for new items. In the previous cycle, 16% of the items had a *p* value below 0.20, compared to 8% in this cycle, which is an improvement. All items had a positive point biserial ranging from 0.03 to 0.53, with a mean of 0.35, which is excellent.

Table 27 Mathematics 2 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	1%
	p Value less than 0.20	8%
	None	90%
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 28 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.

One item was flagged as showing DIF by first language, with this item favouring candidates that spoke English as a first language over those who did not. The item 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 28 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	733085	-0.2084	English	1.87	0.0000

Biology

The DIF results are reported in Table 29 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.

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.

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

Table 29 Biology Items Flagged with C Category DIF

Category	**Item**	***b* Value**	**Group Preferred**	**MH DIF Value**	***p* Value** (significance < 0.001)
Gender	731190	1.9963	Man	2.50	0.0000

Chemistry

The DIF results are reported in Table 30 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. One of these items favoured male candidates, and the remaining item was flagged as preferring both English as a first language candidates and UK White candidates over UK non-White candidates. 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.

Table 30 Chemistry Items Flagged with C Category DIF

Category	**Item**	***b* Value**	**Group Preferred**	**MH DIF Value**	***p* Value** (significance < 0.001)
Gender	731577	1.4335	Man	1.74	0.0000
Ethnicity	731594	-0.5688	UK White	1.71	0.0000
English as a First Language	731594	-0.5688	English	1.51	0.0000

Physics

The DIF results are reported in Table 31 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.

Two items were flagged as showing DIF by Gender, with one item favouring male candidates and the other one favouring female candidates. 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 First Language, UK School Type or UK Ethnicity.

Table 31 Physics Items Flagged with C Category DIF

Category	**Item**	***b* Value**	**Group Preferred**	**MH DIF Value**	***p* Value** (significance < 0.001)
Gender	732777	-2.1609	Man	1.77	0.0000
	732595	-1.0890	Woman	1.54	0.0009

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.