CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Research Matters

Issue 39 / Spring 2025

Proud to be part of the University of Cambridge

Cambridge University Press & Assessment unlocks the potential of millions of people worldwide. Our qualifications, assessments, academic publications and original research spread knowledge, spark enquiry and aid understanding.

Citation

Articles in this publication should be cited using the following example for article 1: Gill, T. (2025). The impact of taking Core Maths on students' higher education outcomes. *Research Matters: A Cambridge University Press & Assessment publication*, 39, 6–25. https://doi.org/10.17863/CAM.116167

Credits

Reviewers: Sylvia Vitello, Matthew Carroll, Emma Walland, Santi Lestari and Simon Child.

Editorial and production management: Lisa Bowett

Additional proofreading: Alison French

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, please contact our team: Research Division, ResearchDivision@cambridge.org

If you need this document in a different format contact us, telling us your name, email address and requirements and we will respond within 15 working days.

Research Matters Issue 39, https://doi.org/10.17863/CAM.116172

All details are correct at the time of publication in March 2025.

Contents

Foreword

Tim Oates, CBE

After the shock of the "Pandemic Years" the educational news has shifted to grand movements in curriculum and assessment — system-level reviews in a substantial number of nations, assessment innovations driven by digital innovation and artificial intelligence (AI). It does feel as though tectonic plates are shifting. But innovation needs to continue at a micro as well as a macro level, driven by a commitment to accumulation of scientific knowledge about learning and measurement. The articles in this edition of *Research Matters* indeed show that research matters — meticulous scrutiny of how innovative approaches in awarding actually work; meticulous examination of pupil work. It may feel like "looking down" when everyone else is looking up, scanning the landscape for massive change, but the hard business of well-designed empirical research needs to continue — it provides the solid base both for innovation in assessment and for curriculum change. A standout feature of exam boards is that they sit on top of massive amounts of data about human performance — exam scripts and the responses to other assessments contain the evidence of what people think, what they know and what they can do. The articles include exemplary approaches to probing these data not just for curiosity-driven purposes but for a means of improving what we do in assessment and learning.

Editorial

Victoria Crisp

Welcome to the spring issue of *Research Matters*. We begin this issue with an article by Tim Gill in which he explores whether taking Core Maths qualifications (at age 16 to 18 years) may have benefits for students during higher education. Specifically, he analyses whether students who studied Core Maths and then went on to begin a degree course with a quantitative element, were less likely to drop out, and more likely to achieve a high degree classification, than those who did not take Core Maths. This follows on from Tim's article in our autumn 2024 issue.

Our second article, by Tom Benton, relates to using comparative judgement (CJ) to support decisions about setting grade boundaries. Awarding processes routinely involve experts' views on the quality of candidate work as one source of evidence and various piloting has explored the potential for CJ to provide this expert input. However, one obstacle is that CJ exercises are time consuming. Tom describes research exploring whether grade boundaries for a whole qualification could be determined based on a CJ exercise for just one exam component rather than needing CJ exercises for each exam component.

In our third article, Santi Lestari considers how certain features of exam question design could plausibly have implications for the accessibility of questions and explores the use of omit rates as a way to monitor for accessibility issues. Santi's analysis used science questions from 44 GCSE exam papers and compared omit rates for questions that required candidates to either create or augment a visual to those for questions that did not. More in-depth analysis also compared omit rates for these item types by other question properties (e.g., position within the exam) and candidate characteristics (e.g., attainment).

Finally in this issue, we have two articles focused on the annotations that learners make when taking exams. Joanna Williamson describes research in which she extracted images of annotations from large samples of GCSE Combined Science and GCSE Mathematics exam scripts in order to derive frequencies of learner annotations, types of annotations, and annotation heat maps (that provide a visualisation of the frequency of annotations in different locations on or around a question). Sylvia Vitello, Abdullah Ali Khan, Heather Mahy, Sarah Hughes and I describe research in which economics learners took a digital multiple-choice exam with access to either scrap paper or a print of the test. Aims included increasing our understanding of annotations and written markings that can support learner thinking during a digital exam and the role of accompanying paper materials. The findings from both these studies have potential implications for functionality within digital testing platforms.

The impact of taking Core Maths on students' higher education outcomes

Tim Gill (Research Division)

Introduction

Core Maths (hereafter, "CM") qualifications were introduced into the curriculum in England in 2014 and were first assessed in 2016. These are qualifications which provide an alternative for students who want to continue with their mathematical education post-16, but do not want to take AS or A Level Maths. They are equivalent in size to half an A Level. In most schools or colleges, students wanting to study CM are required to achieve a pass (grade 4 or higher) at GCSE Maths.

Several different CM qualifications are available, with variation in the focus of the content. For example, some are designed to be taken alongside courses with a statistical element (e.g., A Level Psychology), while others are designed to be taken alongside courses with a more general quantitative element (e.g., A Level Economics).

A small number of previous studies have explored how well the main aims of the CM qualifications (to increase participation in post-16 maths and to help develop students' mathematical knowledge and its application to a range of different areas) have been met. These studies are summarised below.

- Aim 1: Increase participation in post-16 maths:
    - Uptake of CM qualifications has increased over time, from around 3000 entries in 2016 to nearly 13 000 in 2024 (AMSP, 2024).
    - However, the percentage of potential candidates (i.e., those passing GCSE Maths, but not taking A Level Maths) entering the qualification in 2021/22 was only 7 per cent (Royal Society, 2023).
    - There was a significant amount of variation between local authorities in the proportion of schools and colleges offering the subject, i.e., provision was "patchy" (Royal Society, 2023).
- Aim 2: Develop students' mathematical knowledge and its application:
    - In a survey, teachers and students reported that they were positive about CM, particularly its applications to real-world situations (Homer et al., 2020).

- Teachers also believed that CM supported students with other subjects (e.g., A Levels) with mathematical content taken at the same time. However, early analysis found no empirical evidence of improved performance in these subjects (Homer et al., 2020).
- More recent analysis found that in some subjects (with a quantitative element) taken concurrently, CM students performed slightly (but statistically significantly) better than non-CM students (Gill, 2024a).

One of the stated main purposes of CM qualifications was to help "develop students' understanding and application of maths in ways that are valuable for further study and employment across a range of areas" (DfE, 2013, p. 5). This suggests that CM qualifications may help students in their future study (in further or higher education (HE)) in subjects which have some mathematical content, such as sciences, psychology, business, and engineering.

There is some recognition from universities of the benefit of taking CM. Smith (2017) reported that (at the time of writing) 43 universities had shown individual support for CM qualifications, including 23 Russell Group institutions. The Advanced Mathematics Support Programme (AMSP, 2024) lists 10 universities which make lower admissions offers in some subjects to students with a CM qualification. This demonstrates that some universities believe that CM can benefit students in their HE studies.

The main purpose of the research presented here was to investigate whether there is any evidence that taking a CM qualification is helpful to students in terms of HE outcomes (specifically, drop-out rates and degree performance).

The research questions were:

- Are Core Maths students less likely than non-Core Maths students to drop out of HE courses with a quantitative element?
- Is taking Core Maths associated with better degree performance in courses with a quantitative element?

In answering these research questions, we restricted our analysis to HE subjects with some quantitative element, as these are the subjects where taking CM is most likely to be beneficial.

Data and methods

The main source of data for this project was a dataset linking students' records in the National Pupil Database (NPD) and in the Higher Education Statistics Agency (HESA) database. The NPD is administered by the Department for Education (DfE) and includes examination results for all students in all qualifications and subjects in schools and colleges in England, as well as student and school background characteristics such as gender, ethnicity, level of income-related deprivation and school type. The HESA data has information on the students who attend universities in the UK. It includes details of the institution attended, the course subject and level, the degree classification obtained (where applicable) and some additional background characteristics, such as socioeconomic status and level of parental education.

All data was accessed and used in line with the requirements of the organisations that administer these databases. This work was carried out in the Secure Research Service, part of the Office for National Statistics (ONS).

We used the Key Stage 5 (KS5) extract of the NPD for 2017/18 linked to HESA data in 2018/19, 2019/20 and 2020/21. This enabled us to investigate the relationship between taking CM and the probability of dropping out of HE courses with a quantitative element and the probability of achieving a "good" degree (first class or upper second-class) in courses with a quantitative element.

To select the courses with a quantitative element we used the HESA subject classifications, known as the Common Aggregation Hierarchy (CAH).[1] Using the highest level of aggregation, we identified courses from the following classifications as likely to have a quantitative element:

- Biological and sport sciences
- Psychology
- Physical sciences
- Engineering and technology
- Geography, earth and environmental sciences
- Social sciences
- Business and management

Note that subjects in the mathematical sciences group were not included because students taking these subjects would be expected to have A Level Maths and, therefore, are unlikely to have studied Core Maths.

Some students took combined courses where they studied more than one subject. For these students, if more than 50 per cent of the course was in a subject classified as having a quantitative element, then the student was counted as taking a subject with a quantitative element. Otherwise, the student was excluded.

We also excluded students who took AS or A Level Maths. This meant we were able to directly compare students who took CM with those not taking any KS5 maths qualification.

For the analysis of drop-out rates we considered two possible degree start dates (2018/19 and 2019/20). Students who were present in the HESA data (and taking a subject with a quantitative element) in year 1 of their degree but were not present (or were no longer taking a subject with a quantitative element) in year 2 were counted as having dropped out of HE in their first year. This is not a perfect measure, as some of these students may have transferred to a university in a different country or taken a year out (i.e., not dropped out), but we assumed that this was a very small number and would not, therefore, affect the results. We combined data from the two separate start years, so that students who started HE in 2018/19 but were not in the data for 2019/20, and students who started in 2019/20 (i.e., those who deferred a year) but were not in the data for 2020/21, were counted as dropping out.

1 See https://www.hesa.ac.uk/support/documentation/hecos/cah

For the analysis of degree class achieved, we focused on students who were at the end of KS5 in 2017/18 and who completed a degree in 2020/21 (according to the HESA data). This means that the analysis was limited to students who started HE immediately after finishing school and completed their degree in three years. This will therefore exclude any students who took four-year courses, or those who took a year out during their degree.

For both research questions, the initial analysis was descriptive, showing patterns of drop-out and achievement in HE. Then, we carried out logistic regression analyses to fully account for the students' backgrounds when investigating drop-out and attainment for CM and non-CM students.

Regression analysis

For both research questions, logistic regression models were fitted.

The first set of regression models predicted the probability of a student taking a subject with a quantitative element dropping out of HE in their first year.[2] For these models, we used a cross-classified multilevel model, which accounted for two separate hierarchies in the data: students clustered in schools and in HE institutions. For a more detailed description of multilevel logistic regressions see Goldstein (2011). The general form of the model was:

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2ijk} + \ldots + \beta_l x_{lijk} + u_j + u_k$$

where p_{ijk} is the probability of student i from school j attending HE institution k dropping out of HE, x_{1ijk} to x_{lijk} are the independent variables, β_0 to β_l are the regression coefficients, u_j is a random variable at school level and u_k is a random variable at HE institution level.

The second set of models predicted the probability of achieving a first-class degree in a quantitative subject (and separately the probability of achieving at least an upper second-class degree). A cross-classified multilevel model was employed here too, with students nested in schools and in HE institutions. The general form of the model was:

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2ijk} + \ldots + \beta_l x_{lijk} + u_j + u_k$$

where p_{ijk} is the probability of student i from school j and achieving a first (or, separately, at least an upper second) in HE institution k and all other terms are as in the model predicting drop-out.

Analysis was carried out in the R programming language, with the regression models fitted using the *glmer* function from the *lme4* package (Bates et al., 2015).

2 An additional analysis was undertaken predicting the probability of a student dropping out in either year 1 or year 2. The results of this analysis are not presented in this article but are shown in Gill (2024b).

In each regression model, we included contextual variables which were likely to affect the outcome variable. The majority of these variables were taken from the NPD: gender, KS5 attainment, deprivation, ethnic group, first language, special educational needs (SEN) status, total size of qualifications taken at KS5, school type, school gender composition, and school mean KS5 attainment. Other contextual variables were taken from the HESA data: students' socioeconomic classification, their parents' level of education, and the degree subject group. These variables are described in more detail below.

None of these characteristics were directly related to the research questions being addressed, but it was important that they were included in the models because it allowed us to be more confident that any significant effect of taking CM was genuine and not down to differences in the other factors. They were all characteristics which previous research (e.g., Chowdry et al., 2013; Gill, 2017; Vidal Rodeiro, 2019; Gill, 2024c) found to be significant factors in determining the likelihood of drop-out or of degree class achieved.

For the measure of KS5 attainment, we used the students' average KS5 points score. This variable was already in the NPD data and was generated by assigning a points score to each achieved grade[3] and averaging this across all KS5 qualifications (at least equivalent in size to an A Level) taken by a student. The measure, therefore, excluded the grade achieved in CM (for those students who took it), as this is equivalent in size to half an A Level.

For the measure of student deprivation, we used the NPD variable Income Deprivation Affecting Children Index (IDACI), which indicates the proportion of children in a very small geographical area (known as Lower layer Super Output Area or LSOA) living in low-income families.[4] It varies between 0 and 1 and indicates how income-deprived the area is that they live in. As such, it cannot tell us how income-deprived the individual students themselves are but it should be a good proxy for this measure.

Students were grouped in the NPD by their ethnic background: Asian, Black, Chinese, mixed, white, other, and unclassified. Chinese students were in a category of their own in the NPD data, likely because they tend to perform better academically than other Asian students (see, for example, DfE, 2015). Students were also grouped by their first language (English or other).

For students' SEN status, we used the categories in the NPD. These were "No SEN", "SEN, no statement", and "SEN, with statement", with the last of these indicating children requiring the most support.[5]

For the four student characteristics described so far (IDACI score, ethnicity, language, and SEN), around 50 per cent of students had missing data. This

3 For example, a grade A* at A Level was worth 60 points, a grade A worth 50 points, down to a grade E (10 points) and a grade U (0 points). More details on how grades are converted to scores can be found at https://www.gov.uk/government/publications/16-to-19-qualifications-discount-codes-and-point-scores

4 For further information on IDACI calculation, including definitions of children, families, and income deprivation, see Smith et al. (2015).

5 A statement of special educational needs is a legal document which outlines the educational needs of the child and how they will be met by the local education authority.

was because these variables are collected as part of the school census, which independent schools and colleges are not required to complete. As such, this data was mostly missing for students in these school types. Students with missing data for any of these variables were excluded from most of the analysis involving the variables, such as the regression models. However, as including the census variables meant losing a large amount of candidates, we repeated the regression analysis without these variables. This allowed us to include many more candidates, which can help to understand how robust any findings from the first model were.

The student total qualification size variable indicated the total size of the KS5 qualifications taken by each student, measured in A Level equivalents. For example, a student taking three A Levels would have a value of 3. Other qualifications were already assigned an equivalent size in the NPD (e.g., BTECs were equivalent in size to either one, two or three A Levels).

For the analysis by school type, schools were grouped into six categories: comprehensive (including academies and secondary moderns), sixth form colleges, further education (FE) / tertiary colleges, independent schools, selective schools, and other schools. This information was taken from the school type and the admission policy variables in the NPD.

We also categorised schools and colleges by their gender composition (i.e., boys', girls', or mixed). To do this, we calculated the percentage of girls in each school. If this was greater than 95 per cent then the school was categorised as a girls' school, if it was less than 5 per cent it was categorised as a boys' school. Otherwise, it was categorised as a mixed school.

To generate the school KS5 attainment measure (centre KS5 point score), we calculated the average KS5 points score among all students in the school, based on achieved grades.

In the HESA data, students were classified by their socioeconomic status (SES), based on their parents' occupation if they were under 21 or their own occupation if 21 or over. The categories used are standard categories used in the UK census, which run from 1 ("Higher managerial & professional occupations") to 8 ("Never worked & long-term unemployed"), with 9 indicating "not classified" (which includes students).[6]

Students were also classified according to whether at least one of their parents had an HE qualification (e.g., degree, diploma, or certificate of HE) or not.

Finally, the degree subject group was included in some models. This was based on the Common Aggregation Hierarchy (CAH) classification, mentioned earlier in this article.

For each set of regression models, variables which were not statistically significantly different from zero[7] were excluded. A backwards stepwise procedure was used to decide in which order to exclude non-significant variables, starting

6 For a full list of the different categories, see https://www.hesa.ac.uk/collection/c16051/a/sec

7 Statistical significance was determined by the Wald Z-test at the 5 per cent level.

with the variable with the highest *p* value and continuing to remove variables in this way until all were statistically significant. Removing non-significant variables in this way is useful when there are a large number of potential predictor variables, as it makes the final model easier to interpret.

To ensure confidentiality of the data, statistical disclosure controls have been applied to the results (tables and graphs). For example, following HESA disclosure requirements (https://www.hesa.ac.uk/about/regulation/data-protection/rounding-and-suppression-anonymise-statistics) all counts have been rounded up or down to the nearest 5 and counts below 10 and percentages based on counts below 10 have either been suppressed or merged with other counts/percentages.

Results

Are Core Maths students less likely than non-Core Maths students to drop out of HE courses with a quantitative element?

As noted earlier, the definition of drop-out used in the analysis was students who either left HE completely, or those who changed course from a subject with a quantitative element to a non-quantitative subject. Table 1 shows the number of students dropping out in year 1 (Y1) according to this definition (whether or not they took CM in KS5).

Table 1: Drop-out status (Y1) of students starting a quantitative subject

Drop-out status	N students	% students
Did not drop out	65 825	87
Dropped out of HE	4 375	6
Changed to a non-quantitative subject	5 280	7
All who dropped out	9 655	13
All students	**75 480**	**100**

Around 6 per cent of students dropped out completely in year 1 and about 7 per cent changed to a non-quantitative subject. For simplicity, in all further analysis we only look at the combined total drop-outs.

Table 2 presents the numbers and percentages dropping out, by whether CM was taken. This shows that there was very little difference in percentage dropping out for CM (12 per cent) and non-CM students (13 per cent).

Table 2: Drop-out status (Y1) of students starting a quantitative subject, by Core Maths uptake

Taken Core Maths?	N taking quantitative subject	N dropping out	% dropping out
No	73 830	9460	13
Yes	1650	195	12

To look in more detail at drop-out rates, the results of the regression predicting drop-out from a subject with a quantitative element are presented in Table 3. This shows the parameter estimates (with standard errors in brackets). Statistical significance (at the 5 per cent level) is indicated by an asterisk.

For this analysis we fitted three different regression models. In model 1, the statistically significant student and school level variables were included. Model 1a added in significant interaction effects between taking CM and the other predictor variables. Model 2 excluded the census variables, meaning that a much higher number of students were included. We did not try extending model 2 by including interaction effects as the main reason for including this model was to check the robustness of the main model (model 1). Overall, model 1 and model 2 showed similar results for the main effects of interest, indicating that the results in model 1 were not strongly affected by the reduced sample size.

In models 1 and 2, the negative parameter estimates for Core Maths indicated that taking Core Maths was associated with a lower probability of dropping out. The effect is illustrated in Figure 1, which shows the probability for "typical"[8] students with different KS5 points scores (using the results of model 1). However, in model 1 this effect was not statistically significant. This contrasts with model 2, where the parameter estimate was slightly higher and was statistically significant. This difference in statistical significance was partly due to having a much larger number of observations in model 2, leading to a smaller standard error.

Table 3: Regression parameters for models predicting the probability of dropping out (in Y1) of a subject with a significant quantitative element (Model 1 = student level variables; Model 1a = interactions; Model 2 = excluding census variables)

Effect		**Model 1 (n=36 315)**	**Model 1a (n=36 315)**	**Model 2 (n=74 680)**
Intercept		-1.423 (0.058)*	-1.439 (0.058)*	-1.490 (0.058)*
Taken Core Maths	No			
	Yes	-0.155 (0.102)	0.293 (0.174)	-0.198 (0.082)*
Gender	Female			
	Male	-0.149 (0.037)*	-0.148 (0.037)*	-0.172 (0.025)*
KS5 points score		-0.017 (0.002)*	-0.017 (0.002)*	-0.014 (0.001)*
IDACI score		0.604 (0.134)*	0.609 (0.135)*	
Candidate total qualification size		-0.082 (0.034)*	-0.083 (0.034)*	-0.038 (0.017)*

8 For the purpose of exemplification, we define "typical" students as female, attending a comprehensive school, taking a subject in the biological and sport sciences subject group, with parents educated to degree level, and with values of continuous variables equal to the mean. The means for the continuous variables are shown in Table A1 of Appendix A.

Effect		Model 1 (n=36 315)	Model 1a (n=36 315)	Model 2 (n=74 680)
Subject group	Biological & sport sciences			
	Business & management	-0.837 (0.050)*	-0.808 (0.050)*	-0.788 (0.033)*
	Engineering & technology	0.135 (0.086)	0.214 (0.088)*	-0.028 (0.051)
	Geography, earth & environmental sciences	-1.705 (0.169)*	-1.836 (0.183)*	-1.536 (0.121)*
	Physical sciences	-0.704 (0.083)*	-0.701 (0.085)*	-0.651 (0.057)*
	Psychology	-1.341 (0.066)*	-1.316 (0.066)*	-1.166 (0.045)*
	Social sciences	-0.722 (0.050)*	-0.708 (0.050)*	-0.673 (0.034)*
	Combined	-0.398 (0.090)*	-0.404 (0.092)*	-0.246 (0.059)*
Parent educated to degree level	Yes			
	No			0.137 (0.025)*
	Don't know / refused			0.038 (0.037)
School type	Comprehensive / academy			
	6th form college			0.128 (0.042)*
	FE / tertiary college			0.222 (0.038)*
	Independent			-0.048 (0.055)
	Other			0.016 (0.052)
	Selective			-0.021 (0.065)
Taken Core Maths* subject group	Biological & sport sciences			
	Business & management		-1.042 (0.313)*	
	Engineering & technology		-1.213 (0.348)*	
	Geography, earth & environmental sciences		1.646 (0.516)*	
	Physical sciences		-0.148 (0.392)	
	Psychology		-0.977 (0.493)*	
	Social sciences		-0.362 (0.315)	
	Combined		0.162 (0.446)	

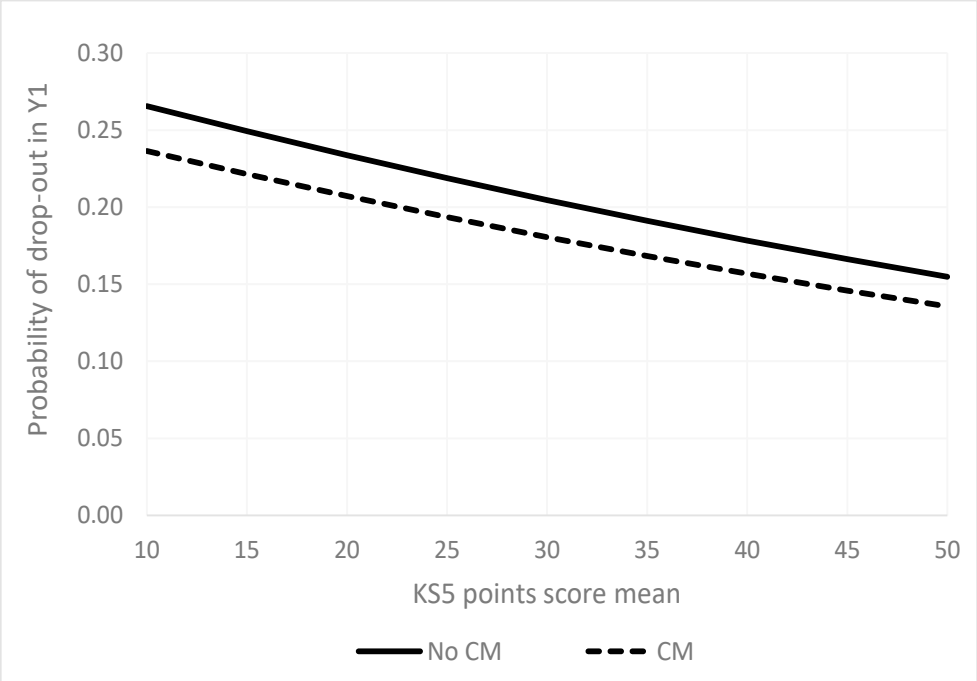


Figure 1: Predicted probabilities of drop-out in year 1, by CM and KS5 mean points score (model 1)

Figure 1 illustrates that the difference in probability according to the model between CM and non-CM students was not large. For example, for students with a mean KS5 points score equal to the mean among all students (33.2, equivalent to one B grade and two C grades at A Level), the probability of dropping out was 0.17 for CM students and 0.20 for non-CM students.

In the model with interactions (model 1a), interpretation of parameter estimates changes. The "main" effect of CM on drop-out rate refers only to the base subject category (biological and sport sciences); from this, we see higher drop-out rates for CM students, but the effect was not statistically significant. The interaction terms (in the bottom rows of the table) show how the effect in that subject differs from the base subject. Several of these were statistically significant, indicating that there was subject-to-subject variation in the effect of CM on drop-out rates. However, from these parameters, we cannot say whether the difference between CM and non-CM students was statistically significant in each subject. Overall, then, the model showed *lower* drop-out rates for CM students in business, engineering, and psychology, but *higher* drop-out rates for CM students in biological sciences, geography, and physical sciences. In combined studies and social sciences, the effect was close to zero. As the size and direction of effects differed between subjects, we illustrate the probabilities of dropping out for CM and non-CM students, by the different subject groups, in Figure 2.

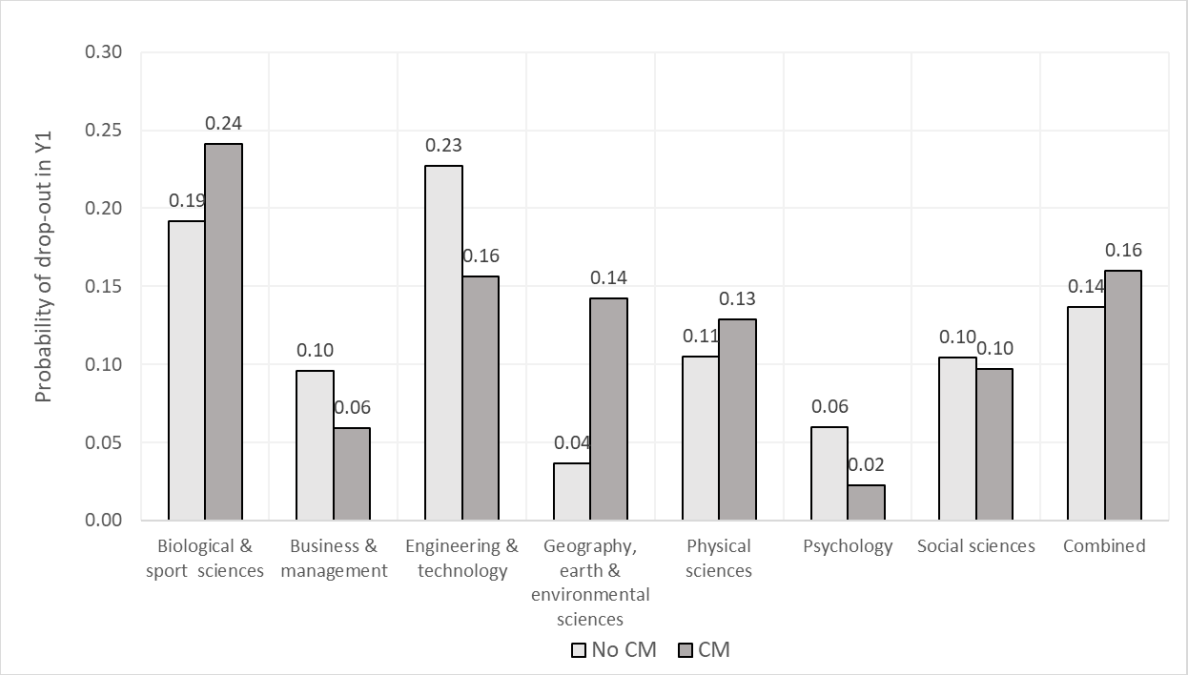


Figure 2: Predicted probabilities of drop-out in year 1, by CM and subject group (model 1a)

Is taking Core Maths associated with better degree performance in courses with a quantitative element?

Achieving a first-class degree

Table 4 shows the overall numbers and percentages of students achieving a first-class degree, by whether they took CM. This shows that CM students were slightly more likely to achieve a first (33 per cent) than non-CM students (29 per cent).

Table 4: First-class degree status, by Core Maths uptake

Taken Core Maths?	N achieving degree in quantitative subject	N achieving a first	% achieving a first
No	31 480	9135	29
Yes	670	220	33

The results of the regression analysis looking at the probability of achieving a first-class degree are presented in Table 5. As before, this shows the parameter estimates (with standard errors in brackets). Statistical significance (at the 5 per cent level) is indicated by an asterisk.

We fitted two different models. In model 1, the statistically significant student and school level variables were included, and model 2 excluded the census variables. We fitted models with interaction effects between CM and the other variables in model 1, but none of these were statistically significant so are not shown here.

Table 5: Regression parameters for models predicting the probability of achieving a first in a subject with a significant quantitative element (Model 1 = student and school level variables; Model 2 = excluding census variables)

Effect		**Model 1 (n=17 230)**	**Model 2 (n=31 795)**
Intercept		-0.216 (0.087)*	-0.291 (0.072)*
Taken Core Maths	No		
	Yes	0.216 (0.111)	0.319 (0.091)*
Gender	Female		
	Male	-0.507 (0.041)*	-0.476 (0.030)*
KS5 points score		0.065 (0.003)*	0.053 (0.002)*
IDACI score		-1.290 (0.175)*	
Candidate total qualification size		0.264 (0.039)*	0.226 (0.024)*
Ethnic group	White		
	Other	-0.321 (0.150)*	
	Asian	-0.182 (0.069)*	
	Black	-0.735 (0.090)*	
	Chinese	-0.052 (0.283)	
	Mixed	-0.299 (0.090)*	
	Unclassified	-0.167 (0.172)	
Language	English		
	Other	-0.303 (0.064)*	
	Unclassified	-0.857 (0.358)*	
Socioeconomic status (SES)	1		
	2	-0.028 (0.051)	-0.066 (0.038)
	3	-0.170 (0.066)*	-0.163 (0.050)*
	4	-0.255 (0.076)*	-0.217 (0.057)*
	5	-0.045 (0.085)	-0.025 (0.066)
	6	-0.219 (0.072)*	-0.309 (0.053)*
	7	-0.234 (0.082)*	-0.328 (0.061)*
	8	-0.108 (0.267)	-0.195 (0.199)
	9	-0.181 (0.067)*	-0.233 (0.048)*
Parents educated to degree level	Yes		
	No		-0.078 (0.031)*
	Don't know / refused		-0.246 (0.047)*

Effect		Model 1 (n=17 230)	Model 2 (n=31 795)
Subject group	Biological & sport sciences		
	Business & management	0.216 (0.060)*	0.183 (0.043)*
	Engineering & technology	0.258 (0.181)	0.316 (0.106)*
	Geography, earth & environmental sciences	-0.207 (0.081)*	-0.090 (0.062)
	Physical sciences	0.557 (0.138)*	0.371 (0.095)*
	Psychology	-0.467 (0.063)*	-0.379 (0.048)*
	Social sciences	-0.338 (0.058)*	-0.271 (0.043)*
	Combined	-0.348 (0.118)*	-0.231 (0.089)*
School type	Comprehensive / academy		
	6th form college	0.088 (0.179)	-0.278 (0.055)*
	FE / tertiary college	0.761 (0.582)	-0.673 (0.054)*
	Independent	0.036 (1.331)	-0.046 (0.067)
	Other	-0.184 (0.062)*	-0.225 (0.061)*
	Selective	0.225 (0.070)*	0.205 (0.070)*
Centre KS5 points score		-0.029 (0.005)*	-0.012 (0.004)*

The results show that there was a positive effect of taking CM on the probability of achieving a first in a quantitative subject. However, this difference was not statistically significant in the main model (model 1). The size of the effect is illustrated in Figure 3, which shows the probabilities for "typical"[9] CM and non-CM students at different levels of KS5 mean points score. For example, at the mean value of KS5 points score mean (35.1) CM students had a probability of a first of 0.50, compared with 0.45 for non-CM students.

Comparing model 1 with model 2, the effect of excluding the census variables and increasing sample size on the parameter estimates was small. However, there were some differences in the statistical significance of these estimates, with the parameter estimate for taking CM not significant in model 1 and significant in model 2. This was due in part to having a much larger number of observations in model 2, leading to a smaller standard error.

9 We define "typical" students in this case to be female, attending a comprehensive school, taking a course in the biological sciences subject group, with parents educated to degree level, in socioeconomic classification group 1, and with values of continuous variables equal to the mean. The means for the continuous variables are shown in Table A2 of Appendix A.

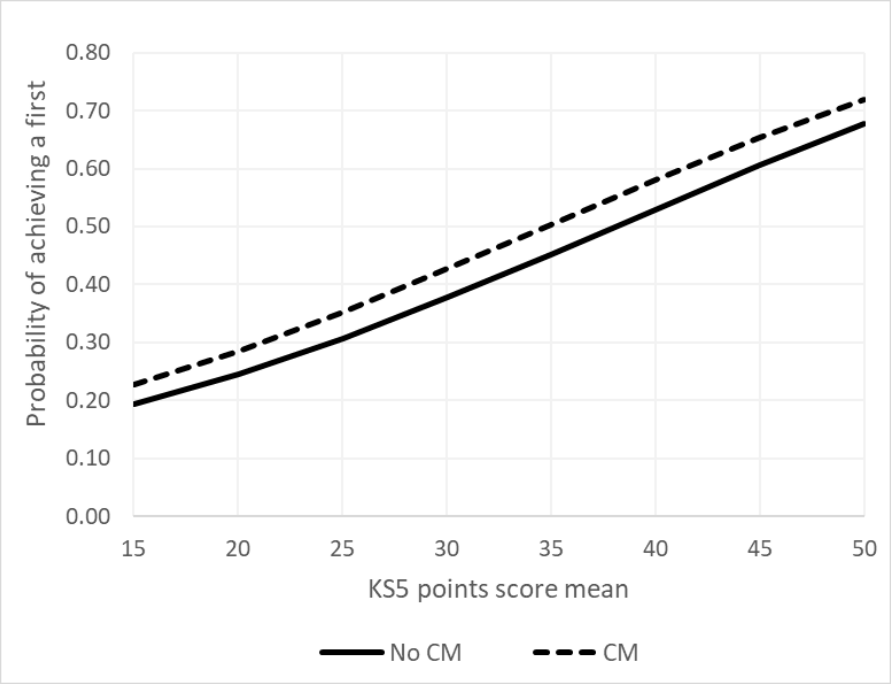


Figure 3: Predicted probabilities of achieving a first, by CM uptake and KS5 mean points score (model 1)

Achieving at least an upper second-class degree

Table 6 shows the overall numbers and percentages of students achieving an upper second-class degree or higher, by CM uptake. This shows that CM students were slightly more likely to achieve at least an upper second (87 per cent) than non-CM students (84 per cent).

Table 6: Upper second-class degree (or higher) status, by Core Maths uptake

Taken Core Maths?	N achieving degree in quantitative subject	N achieving at least an upper second	% achieving at least an upper second
No	31 480	26 490	84
Yes	670	580	87

The results of the regression models looking at the probability of achieving at least an upper second-class degree are shown in Table 7. In model 1, the significant student and school level variables were included, and model 2 excluded the census variables. We again fitted models with interaction effects between taking CM and the other variables in model 1, but none of these were statistically significant, so we do not show the results.

Table 7: Regression parameters for models predicting the probability of achieving at least an upper second in a subject with a significant quantitative element (Model 1 = student and school level variables; Model 2 = excluding census variables, due to missing data)

Effect		**Model 1 (n=17 230)**	**Model 2 (n=31 795)**
Intercept		2.383 (0.100)*	2.220 (0.077)*
Taken Core Maths	No		
	Yes	0.426 (0.160)*	0.350 (0.122)*
Gender	Female		
	Male	-0.560 (0.053)*	-0.522 (0.036)*
KS5 points score		0.056 (0.003)*	0.040 (0.002)*
IDACI score		-1.547 (0.206)*	
Candidate total qualification size		0.292 (0.057)*	0.226 (0.033)*
Ethnic group	White		
	Other	-0.293 (0.167)	
	Asian	-0.197 (0.085)*	
	Black	-0.748 (0.089)*	
	Chinese	0.910 (0.539)	
	Mixed	-0.172 (0.111)	
	Unclassified	0.051 (0.227)	
Language	English		
	Other	-0.164 (0.074)*	
	Unclassified	0.084 (0.367)	
Socioeconomic status (SES)	1		
	2	0.006 (0.077)	-0.120 (0.053)*
	3	-0.266 (0.089)*	-0.350 (0.062)*
	4	-0.177 (0.101)	-0.216 (0.071)*
	5	-0.123 (0.119)	-0.192 (0.084)*
	6	-0.200 (0.093)*	-0.435 (0.064)*
	7	-0.299 (0.103)*	-0.424 (0.072)*
	8	-0.536 (0.287)	-0.373 (0.212)
	9	-0.208 (0.091)*	-0.374 (0.061)*
Parents educated to degree level	Yes		
	No		-0.002 (0.038)
	Don't know / refused		-0.162 (0.055)*

Effect		Model 1 (n=17 230)	Model 2 (n=31 795)
Subject group	Biological & sport sciences		
	Business & management	0.552 (0.078)*	0.527 (0.051)*
	Engineering & technology	0.099 (0.220)	0.329 (0.122)*
	Geography, earth & environmental sciences	0.287 (0.121)*	0.442 (0.090)*
	Physical sciences	0.459 (0.207)*	0.392 (0.122)*
	Psychology	0.200 (0.085)*	0.326 (0.060)*
	Social sciences	0.062 (0.074)	0.100 (0.051)*
	Other	0.089 (0.152)	0.189 (0.103)
School type	Comprehensive / academy		
	6th form college	0.131 (0.245)	-0.379 (0.064)*
	FE / tertiary college	0.310 (0.694)	-0.733 (0.059)*
	Independent	-3.291 (1.340)*	0.185 (0.083)*
	Other	-0.126 (0.075)	-0.187 (0.074)*
	Selective	0.462 (0.112)*	0.463 (0.106)*
Centre KS5 points score		-0.018 (0.007)*	

These results show a significant and positive effect of taking CM on the probability of achieving at least an upper second. This is illustrated in Figure 4, which shows the probabilities for "typical"[10] students with different levels of KS5 mean points score (using the results of model 1).

The size of the effect was not large: at the mean value of KS5 points score mean (35.1) CM students had a probability of a first of 0.94, compared with 0.92 for non-CM students. For higher values of the KS5 points score mean the probabilities for CM and non-CM students were even closer.

There were mostly only small differences between the parameter estimates in model 1 (including census variables) and model 2 (excluding census variables). In particular, the estimate for taking CM fell from 0.426 to 0.350. As there was no change to the statistical significance of this estimate, the finding that taking CM was beneficial was unchanged.

10 We define "typical" students in this case to be female, white, with English as their first language, attending a comprehensive school, taking a course in the biological and sport sciences subject group, with parents educated to degree level, in socioeconomic classification group 1, and with values of continuous variables equal to the mean. The means for the continuous variables are shown in Table A3 of Appendix A.

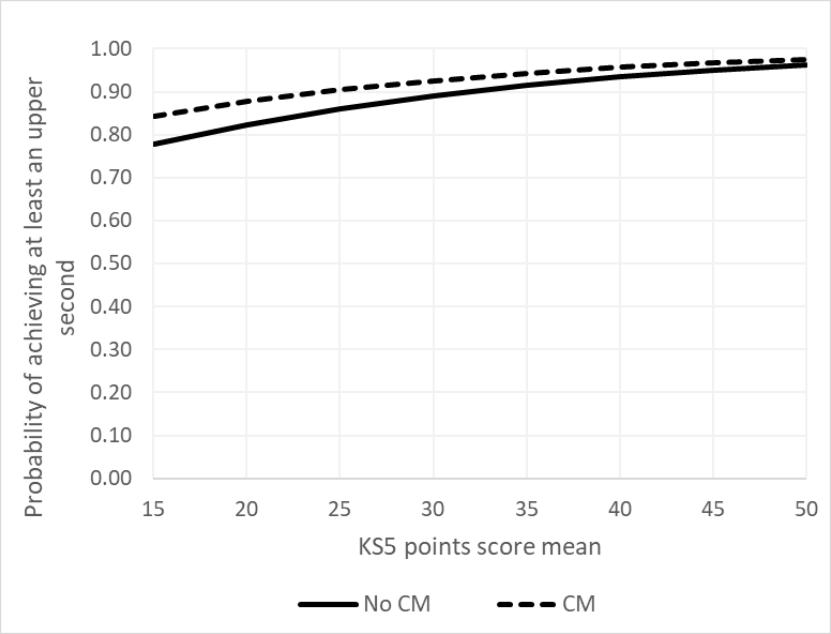


Figure 4: Predicted probabilities of achieving at least an upper second, by CM uptake and KS5 mean points score (model 1)

Conclusions

The main purpose of the analysis presented in this article was to investigate whether CM was beneficial, in terms of HE outcomes, for those students taking it.

The results presented here were part of a more comprehensive analysis into the potential benefits of taking CM (see Gill, 2024b). One of the main findings from that analysis (but not shown in this article) was that CM students were significantly more likely to progress to HE in a subject with a quantitative element (probability of 0.49 for a typical CM student compared to 0.39 for a typical non-CM student). This was not a surprising finding as many students will have taken the qualification in the expectation of studying further in a quantitative subject.

The results presented in this article focused on outcomes at HE, specifically whether students taking Core Maths were less likely to drop out of HE, and more likely to achieve a good HE degree in subjects with a quantitative element, than those not taking the qualification.

In terms of drop-out rates, descriptive statistics indicated that CM students were slightly less likely to drop out than non-CM students. Statistical models showed somewhat variable effects. In the models with a smaller sample (but including census variables), the effect was subject-dependent, with negative effects on drop-out rate seen in business and management, engineering and technology, and psychology courses. Somewhat surprisingly, there were *positive* effects (i.e., a greater drop-out rate for CM students) in biological and sport sciences, and geography, earth and environmental sciences. Other subject groups showed only very small effects. When the whole sample was included (but census variables were excluded) there was an overall negative effect of taking CM on dropping out: i.e., students that took CM were significantly less likely to drop out. Overall, then, it seems that taking CM can be associated with reduced risk of dropping out of HE, but not across all subjects.

In terms of the probability of students gaining an upper second-class degree, CM had a significant, positive impact regardless of the model and sample used. The effect was small (0.94 for CM students, 0.92 for non-CM students), but this might be because such a high proportion of students achieved an upper second-class degree anyway. In terms of the probability of students gaining a first-class degree, the models again indicated a positive effect of taking CM, but this was not statistically significant in the main model. However, in the model with the larger sample this was significant (probability of 0.51 for CM students, 0.43 for non-CM students).

Perhaps surprisingly, there was no evidence of differences in the effect of taking CM on degree outcomes for the different subject groups (i.e., no significant interaction effect between CM and subject group). This may be related to using high-level subject grouping in the regression analysis. Using finer subject classifications instead might have identified significant differences between subjects in the effect of taking CM on degree outcomes, perhaps due to their differences in mathematical content. Alternatively, the issue may be that our analysis is limited by the fairly small numbers of CM students taking some individual subjects within each subject group.

Taken together, these findings suggest that taking CM may be beneficial to students taking a quantitative subject at HE. When we included the full sample of students, those taking CM were significantly less likely to drop out and significantly more likely to achieve a good degree. When a more limited sample was included, permitting additional contextual variables to be included, effects were similar but were not always statistically significant, or appeared to vary between subjects. Nevertheless, the overall results are encouraging and suggest that CM can help in the way it is intended to. These findings should encourage more universities to follow the policy of making reduced offers to students with CM or to welcome its addition to students' programmes of study.

Finally, there was one notable limitation with this research: that association does not mean causation. There may be other reasons why CM students were less likely to drop out and more likely to achieve a good degree that were not directly related to taking CM. For instance, it may be that students taking CM were more motivated to do well academically than non-CM students and it was this that led to better outcomes at HE, rather than taking CM per se. It would be interesting to undertake further research into this, by speaking with students in HE (across a range of different subject areas) who took CM, to find out their motivation for taking CM and whether they believed it had helped them with their HE studies.

Acknowledgements

This work was carried out in the Secure Research Service, part of the Office for National Statistics (ONS). It contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

References

AMSP. (2024). *Level 3 maths update 2024-25*. Advanced Mathematics Support Programme.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Chowdry, H., Crawford, C., Dearden, L., Goodman, A., & Vignoles, A. (2013). Widening participation in higher education: Analysis using linked administrative data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(2), 431–457.

DfE. (2013). *Introduction of 16 to 18 core maths qualifications. Policy statement.* Department for Education.

DfE. (2015). *Statistical First Release: GCSE and equivalent attainment by pupil characteristics, 2013 to 2014 (Revised)*. Department for Education.

Gill, T. (2017). Preparing students for university study: A statistical comparison of different post-16 qualifications. *Research Papers in Education*, *33*(3), 301–319.

Gill, T. (2024a). Core Maths: Who takes it, what do they take it with, and does it improve performance in other subjects? *Research Matters: A Cambridge University Press & Assessment publication, 38*, 48–65.

Gill, T. (2024b). *Is Core Maths fulfilling its aim? Impact on higher education outcomes*. Cambridge University Press & Assessment.

Gill, T. (2024c). The extended project qualification in England: Does it provide good preparation for higher education? *Oxford Review of Education*.

Goldstein, H. (2011). *Multilevel statistical models (4th edition*). John Wiley & Sons.

Homer, M., Mathieson, R., Tasara, I., & Banner, M. L. (2020). *The early take-up of Core Maths: successes and challenges*. University of Leeds.

Royal Society. (2023, December 13). *Why Core Maths?*

Smith, A. (2017). *Report of Professor Sir Adrian Smith's review of post-16 mathematics*. Department for Education.

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., & Plunkett, E. (2015). *The English Indices of Deprivation 2015 Technical report*. Department for Communities and Local Government.

Vidal Rodeiro, C. L. (2019). The impact of A Level subject choice and students' background characteristics on Higher Education participation. *Research Matters: A Cambridge Assessment publication*, 28, 17–26.

Appendix A

Table A1: Mean values of continuous variables used in regression models (probability of drop-out year 1)

Variable	N students	Mean
KS5 points score	75 095	33.23
IDACI score	36 440	0.17
Candidate total qualification size	75 480	3.19
Centre KS5 points score	75 125	32.03

Table A2: Mean values of continuous variables used in regression models (probability of first)

Variable	N students	Mean
KS5 points score	31 980	35.09
Candidate total qualification size	32 150	3.30
Centre KS5 points score	31 990	32.26

Table A3: Mean values of continuous variables used in regression models (probability of at least an upper second)

Variable	N students	Mean
KS5 points score	31 980	35.09
IDACI score	17 315	0.17
Candidate total qualification size	32 150	3.30
Centre KS5 points score	31 990	32.26

Is one comparative judgement exercise for one exam paper sufficient to set qualification-level grade boundaries?

Tom Benton (Research Division)

Introduction

In high-stakes qualifications such as GCSEs and A Levels in England, grade boundaries refer to the minimum score a candidate requires to achieve a given grade. Historically, the way in which these have been determined has always included an element of human expert judgement (Benton & Elliott, 2015). Most commonly this has been in the form of expert examiners in the relevant subjects reviewing examination scripts with scores close to suggested grade boundaries and indicating whether or not they feel they are of sufficient quality to be awarded the focus grade. However, for some time, a number of authors have argued that comparative judgement (CJ) may provide a more effective means of incorporating inputs from subject experts into the process of determining grade boundaries (e.g., Bramley, 2007; Curcin et al., 2019). The awarding body OCR conducted a substantial programme of research investigating this technique, which is described in Benton et al. (2022).

In the context of awarding, a typical CJ exercise involves judges (usually examiners) comparing many pairs of scripts. Each pair consists of one script from the current exam series and one from a previous series. Note that the scripts from different series will consist of responses to different sets of questions. Allowing for this fact, judges must decide which script in each pair displays superior overall performance in the subject. Several hundred such comparisons are completed. The decisions from expert judges are then analysed with respect to the scores that were awarded to the different scripts with the aim of identifying pairs of scores in each series that, based on expert judgement, display equivalent levels of performance. For the CJ exercises in this article, we define two scores as equivalent if, when scripts with the two scores from the two respective series are compared, judges are equally likely to select either one as superior. For further details of the method see Benton (2021) or Benton et al. (2020). A slightly different methodology for using CJ in awarding is described by Curcin et al. (2019).

At present, if CJ is used in awarding, a separate exercise is carried out for each component of which a qualification is comprised. For example, suppose

a qualification consists of two exam papers. We typically wish to set grade boundaries on the qualification that reflect performance standards at some point in the recent past (often within the prior year). In order to do this, we would complete one CJ exercise for paper 1 to identify grade boundaries that reflect equivalent levels of performance to grade boundaries on the previous version of this same paper in a previous exam series. We would also complete a separate exercise for paper 2 to identify the appropriate boundaries on this paper. Finally, boundary scores at each grade would be added together across papers to set grade boundaries for the whole qualification. For example, if 10 marks were needed to achieve a grade E on paper 1, and 12 marks were needed for a grade E on paper 2, then we would know that, for the qualification as a whole, the grade E boundary should be 22.

As noted by Benton et al. (2022), the use of CJ in awarding requires substantial time from expert judges. With this in mind, it is of interest to explore ways of making the application of CJ in awarding more efficient. One possibility for increased efficiency would be to only conduct a single CJ exercise for a single component. In theory, the outputs of the CJ exercise provide a complete mapping of scores on the previous version of the component to equivalent scores on the current version. Therefore, once this mapping is applied, we have access to a measure of the abilities of the students taking the alternative versions of the qualification (from different exam series) on the same scale — that is, essentially an anchor test. We might then use the scores on this anchor component to perform complete test equating of scores on the different qualification versions using any equating method we choose from those applicable to a non-equivalent anchor test (NEAT) design (Kolen & Brennan, 2004).

Although it is clear that the above approach would require less time from expert judges than completing a CJ exercise for every component, it is not known whether it would provide accurate results. In this article, we explore this issue in more detail. Firstly, we examine the consistency of evidence from different individual components regarding changes in cohort ability between series. This is of interest as, if we are to rely on evidence from a single component, it is important that changes in performance in one component are indicative of changes in performance on the qualification as a whole.

Having done this, we then explore in detail whether the results from separate CJ exercises on different components necessarily lead to the same grade boundaries at qualification level. Furthermore, we evaluate the impact of using CJ from a single component upon the standard errors of estimated grade boundaries compared to the current approach of combining several separate CJ exercises. That is, even if we assume that a single component is sufficient to indicate how the performance of a cohort has changed, what is the impact of using just one component on the precision of the technique?

Data

This research re-analysed data from three qualifications from Benton et al. (2022) where all of the components comprising a qualification were included in separate

CJ exercises. Further details of the CJ exercises conducted for each of the three qualifications, referred to as "projects" for the purposes of this article, are given in Table 1. The projects all relate to qualifications that were awarded using CJ in autumn 2020. In each of these cases, a "simplified pairs" CJ approach (Benton, 2021; Benton et al., 2020) was taken. For all three of the projects, the aim was to carry forward performance standards from June 2019 to set grade boundaries in autumn 2020.

Table 1 shows the amount of available data in each of the CJ exercises. Note that, due to the unusual nature of the autumn 2020 exam series,[1] very few scripts were available. As such, fewer scripts from this series than from June 2019 were included in the study. Because of this, each script from June 2019 was included in a single paired comparison whereas scripts from the autumn series were often included in multiple pairs.

Table 1: Details of the CJ exercises

Project ID	Qualification	Subject	Paper	Max. score	N scripts		N judges	N pairs
					June 2019	Autumn 2020		
1	A Level	English Literature	1	60	466	91	6	466
			2	60	414	97	5	414
2	A Level	Psychology	1	90	498	66	6	498
			2	105	500	53	6	500
			3	105	500	51	6	500
3	GCSE	English Language	1	80	350	291	6	350
			2	80	350	345	6	350

As well as the data from the CJ exercises, data on the scores achieved in each exam paper by all those who took the qualifications was used for some analyses.[2]

Relationship between scores in different components

Intrinsically, any justification for using a single component as an anchor for an entire qualification is dependent upon the different components measuring broadly the same abilities. Of course, different components tend to cover different topic areas within a subject. However, we would hope that they all rely on broadly the same underlying set of knowledge and skills. Some evidence of this can be provided by looking at the correlations between scores on different components for the full cohort of candidates that took each qualification. These are shown in Table 2. As can be seen, within each series, scores for all of the components within a qualification display fairly strong correlations with one another. These

1 The autumn 2020 exam series was specially arranged to allow students who were unhappy with the grades they were awarded by their school during the pandemic to sit formal exams. Autumn exams are not normally made available for GCSEs or A Levels, except for GCSEs in English Language and Mathematics.

2 The data used in this research was collected as part of the operational marking and processing of candidates' examination scripts. Data has been stored and used in line with Cambridge University Press & Assessment's Data Privacy notice (https://www.cambridge.org/legal/candidate-privacy-notice).

correlations provided initial evidence that it was reasonable to conduct the analyses reported in this article in order to explore the potential of using a CJ exercise on just one component to inform awarding decisions.

Table 2 also shows the total number of candidates in each series. From this we can see that the amount of available data in the autumn series was quite low for two of the qualifications.

Table 2: Correlations between components within each qualification in each exam series

Project ID	Qualification	Subject	Papers	Correlation in...		N candidates	
				June 2019	Autumn 2020	June 2019	Autumn 2020
1	A Level	English Literature	1 and 2[3]	0.58	0.62	9677	119
2	A Level	Psychology	1 and 2	0.69	0.73	5567	70
			1 and 3	0.66	0.69		
			2 and 3	0.72	0.78		
3	GCSE	English Language	1 and 2	0.81	0.84	13 199	496

Summary of CJ results for individual components

All of the results in subsequent sections are derived from the mappings from June 2019 to autumn 2020 scores that were identified for each individual component using CJ. These mappings are displayed in Figure 1. For each score on each June 2019 paper, the solid black line in each chart shows the score on the autumn 2020 paper that was estimated to be equivalent. An "equivalent" score means one where expert judges would be equally likely to consider a script with this score better than, or worse than, a June 2019 script with the associated score on the x-axis. Figure 1 also shows 95 per cent confidence intervals (CI) as dotted lines, calculated using the method described in Benton et al. (2020). A faint grey line of equality is included in each chart to aid interpretation.

For the English Literature papers, Figure 1 shows the papers in autumn 2020 were perhaps slightly easier than those in June 2019 at the top end of the score distribution — that is, the equivalent scores are significantly above the line of equality. However, at the lower end, the autumn 2020 exams were perhaps easier. Having said this, due to a lack of scripts with low scores in autumn 2020, the confidence intervals are very wide at the lower end.

Figure 1 suggests that, for Psychology, paper 1 was of very similar difficulty in autumn 2020 and June 2019, paper 2 was slightly harder in autumn 2020, and paper 3 was slightly easier. Finally, for English Language, Figure 1 suggests that paper 1 was of similar difficulty in autumn 2020 and June 2019 but that paper 2 was slightly harder.

3 In an ordinary series, A Level English Literature also includes an additional component of non-examined assessment. However, due to the unusual nature of the autumn 2020 series, the qualification was awarded without this element on this occasion.

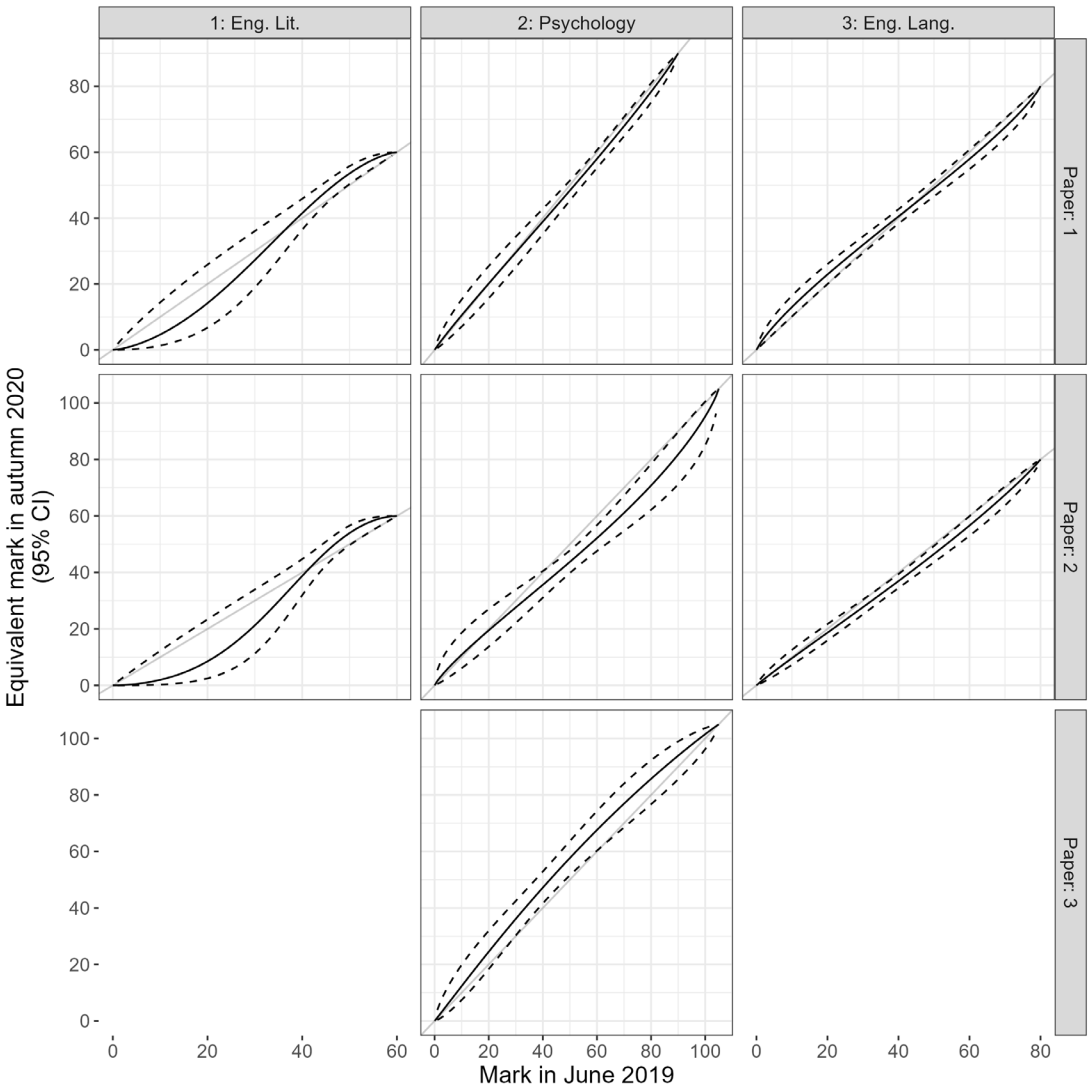


Figure 1: Mapping for each component of each subject based on analysis of CJ data.

Table 3 shows the same results in tabular form for the key grade boundaries[4] on each June 2019 paper. For ease of presentation, equivalent scores on the autumn 2020 papers have been rounded to whole numbers and confidence intervals are presented in terms of each score being correct plus or minus a particular (rounded) value. Across the grades and papers in Table 3, the majority of equivalent scores were identified with a precision of no worse than plus or minus 3 marks. However, some were larger; for example, the widest confidence interval (English Literature, paper 2, grade E) had a precision of plus or minus 12 marks.

4 Key grades are those where regulations require awarding organisations to make explicit decisions about boundaries. The key grades are grades E, A and A* for A Level, and grades 1, 4, 7 and 9 for GCSEs. All other grade boundaries are usually set by linear interpolation between these.

Table 3: Summary of score mappings and confidence intervals based upon CJ for individual components at key grades

Project ID	Subject	Grade	Paper 1		Paper 2		Paper 3	
			June 2019 boundary	Autumn 2020 equivalent	June 2019 boundary	Autumn 2020 equivalent	June 2019 boundary	Autumn 2020 equivalent
1	English Literature	E	26	22 +/- 9	27	17 +/- 12	-	-
		A	53	56 +/- 2	54	58 +/- 3	-	-
		A*	56	58 +/- 2	57	59 +/- 1	-	-
2	Psychology	E	29	29 +/- 5	22	21 +/- 7	32	38 +/- 6
		A	69	67 +/- 3	62	54 +/- 5	69	76 +/- 8
		A*	75	73 +/- 3	71	62 +/- 6	77	83 +/- 8
3	English Language	1	8	11 +/- 3	8	8 +/- 2	-	-
		4	34	35 +/- 2	36	33 +/- 2	-	-
		7	53	52 +/- 3	54	50 +/- 3	-	-
		9	65	63 +/- 3	66	63 +/- 4	-	-

Evidence about cohort ability from different components

Using the mappings shown in Figure 1, it was possible to transform the scores for all candidates who took each paper in June 2019 to the equivalent scores on the autumn 2020 papers. Having done this, we can compare the performance of candidates on each paper between series.

This comparison is shown in the form of boxplots in Figure 2. The top and bottom of each box in the figure represents the 25th and 75th percentiles of the total scores on each paper. The central lines within each box represent the medians and the whiskers represent the range of scores that were seen excluding outliers. Note that the scores from June 2019 have been transformed using the results from the CJ studies (Figure 1) so that, theoretically, if the CJ exercise has worked as intended, scores in the different series are directly comparable as if candidates had taken the exact same versions of each assessment. Therefore, differences in performance in Figure 2 potentially indicate differences in the subject ability of the candidates entering in the different series.

From Figure 2 we can see that, in every paper, the performance of candidates was stronger in June 2019 than autumn 2020. This is unsurprising since the autumn series was mainly intended for candidates who had not achieved the grades they wanted during summer 2020. As such, it is expected that autumn 2020 would attract entries from weaker candidates.

For English Language the difference in the performance of candidates in the two series is very consistent across each paper. Specifically, in each paper, the median performance in June 2019 is just slightly above the 75th percentile of performance in autumn 2020. For English Literature differences are also fairly consistent in that, for each paper, the median performance in June 2019 is between the median and 75th percentile of performance in autumn 2020.

However, for Psychology the three papers show very different patterns. Based on paper 1, candidates in June 2019 were only slightly stronger than those in autumn 2020. Differences on paper 2 appear a little larger and on paper 3 the difference appears huge with the median performance in June 2019 well above the 75th percentile of performance in autumn 2020.

Overall, the results here show that different components can potentially lead to different conclusions about the relative strengths of groups of candidates. For this reason, it is good that we have direct evidence on relative performance levels in all of them rather than relying on a single component. We explore the impact of these differences on grading the qualification overall in the next section.

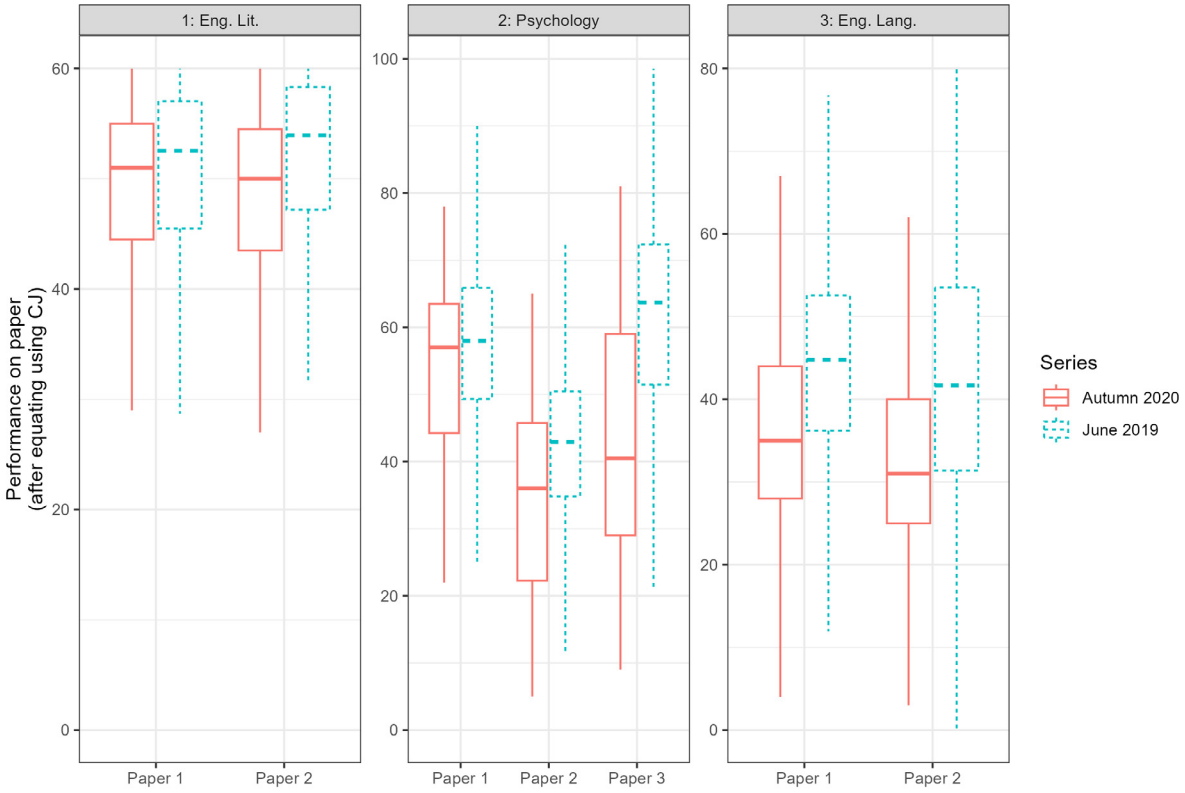


Figure 2: Boxplots showing the distribution of the scores on each paper in June 2019 and autumn 2020. Scores from June 2019 have been transformed using the results from the CJ analysis (Figure 1) so that they are, theoretically, directly comparable with those from autumn 2020.

Impact of using a single component on overall grade boundaries

For each subject, the results from the CJ exercises were used to identify qualification grade boundaries for autumn 2020 equivalent to grade boundaries from June 2019. Note that, due to the impact of the pandemic, and discussions over how the standards required in autumn 2020 exams should relate to those from grades awarded purely via teacher assessment in summer 2020, these do not match the June 2019 boundaries that were used in the actual awarding exercises for these qualifications. However, for the purposes of this research, we will imagine that standards from June 2019 were carried forward to autumn 2020 in a straightforward manner.

Overall grade boundaries for autumn 2020 for each subject were calculated by adding the relevant grade boundaries from different components together — that is, the estimated boundaries in Table 3. Standard errors of estimated grade boundaries at whole qualification level were calculated by taking the square root of the sum of the squared standard errors across the constituent components.

Next, we generated grade boundaries for each whole qualification but based upon each separate available component on its own. The following process was used to achieve this:

1. Collate a dataset for all candidates in either June 2019 or autumn 2020 (i.e., not just those candidates in the CJ exercise) with scores on all constituent components with the following pieces of data:
    a. Each student's score on the component of interest.
    b. Each student's total score on the entire qualification.
2. For each student in the June 2019 data from step 1, replace their component score on the component of interest with the equivalent score from the autumn series as defined in Figure 1. Note that their total score on the whole qualification should not be adjusted.
3. Adjusted component scores from June 2019 and unadjusted scores on the same component in autumn 2020 are now treated as if they are interchangeable. As such, these two sets of scores are treated as an anchor test to allow whole qualification scores from June 2019 to be equated to equivalent scores at whole qualification level in autumn 2020.
4. Use the results from step 4 to identify equivalent values in autumn 2020 to the June 2019 grade boundaries at whole qualification level.

For step 3, chained equipercentile equating was used (Kolen & Brennan, 2004, p. 145). Briefly, equipercentile equating means that, where two assessments have been taken by the same group of candidates, equivalent scores are identified as those that are at the same percentile in the distribution. The "chained" element means this was applied in two steps – first to map grade boundaries on the whole qualification in June 2019 to equivalent points on the anchor, and then to map these anchor points to appropriate positions in the autumn series.

Note that loglinear models were used to smooth the score distributions before the chained equipercentile procedure was applied. Specifically, the empirical score distributions were replaced with smooth versions that retained the mean, standard deviation, skewness and kurtosis of scores on each paper in each series. This method was necessary to address the small sample sizes in the autumn 2020 series. Without smoothing, the large gaps between the scores that actually occur in the data could manifest themselves in some unusual results. In a normal summer exam series, most subjects have entries from rather larger numbers of candidates and such issues do not occur. Thus, the use of smoothing helps ensure the results here are more indicative of what might happen in practice more widely.

In addition to producing estimates of grade boundaries at whole qualification level based upon each separate component, it was of interest to produce standard errors from using a single CJ exercise. These standard errors are intended to show how the precision of a CJ exercise for a single component (Table 3, Figure 1) manifests itself when applied to setting grade boundaries for an entire qualification. The standard errors do not incorporate the uncertainty in the equating process itself (step 3). In most practical situations, with larger sample sizes, this source of uncertainty would be trivial compared to the uncertainty stemming from the CJ exercise in any case.

Standard errors were estimated as follows:

A. The mappings in Figure 1 are based on the coefficients from a logistic regression (see Benton et al., 2020, for further details). Rather than using point estimates of these coefficients, these were sampled from a multivariate normal distribution with a mean at the estimated coefficients and using the variance-covariance matrix of the model parameters.
B. Apply steps 1 to 4 (above) based on a mapping derived using the logistic regression coefficients sampled in A to derive a fresh estimate of the qualification grade boundary.
C. Repeat steps A and B 500 times and use the standard deviations of the estimated boundaries across these repetitions as the standard errors.

For each boundary, 95 per cent confidence intervals were calculated by the usual approximation of multiplying the standard errors by 1.96.

The results of this analysis are shown in Table 4. Table 4 shows the estimated grade boundaries and confidence intervals at whole qualification level based on CJ evidence from each individual component only and also (the final column) based on all of the CJ evidence across all components combined. For ease of reading, all estimated boundaries and confidence interval widths have been rounded to whole numbers.

Note that the estimates from using all components need not always fall between the estimates from individual components.[5] Also note that, due to the impact of rounding, the estimated overall qualification boundaries in the final column may not perfectly match the sum of the estimated values for each paper shown in Table 3.

5 This reflects the fact that, in ordinary chained equating, if we had two possible anchor tests A1 and A2, chained equating using the sum of both anchor tests as the anchor would not give the same result as taking the average of analyses using each anchor test separately. This is because the summed anchor test will have different reliability as well as a differently shaped distribution to either of the individual anchors.

Table 4: Estimated qualification-level grade boundaries in autumn 2020 and standard errors based on different individual components

Project ID	Subject	Grade	June 2019 boundary	Estimated qualification-level grade boundary from source component (paper) and 95% confidence interval											
				Paper 1			Paper 2			Paper 3			All (Original)		
1	English Literature (Max. score =120)	E	53	43	+/-	20	40	+/-	11		-		39	+/-	15
		A	107	113	+/-	4	114	+/-	4		-		114	+/-	3
		A*	113	117	+/-	2	118	+/-	2		-		118	+/-	2
2	Psychology (Max. score =300)	E	83	73	+/-	9	87	+/-	21	88	+/-	17	88	+/-	10
		A	200	190	+/-	11	186	+/-	13	206	+/-	17	197	+/-	10
		A*	223	215	+/-	10	207	+/-	17	223	+/-	16	218	+/-	11
3	English Language (Max. score =160)	1	16	15	+/-	5	21	+/-	4		-		19	+/-	4
		4	70	68	+/-	5	67	+/-	5		-		69	+/-	3
		7	107	100	+/-	6	104	+/-	6		-		102	+/-	4
		9	131	123	+/-	7	131	+/-	6		-		126	+/-	5

The same information in Table 4 is displayed in a different way in Figure 3. Figure 3 displays estimated grade boundaries for autumn 2020 in terms of how far they moved from June 2019. It also shows 84 per cent confidence intervals for each estimated change as, according to Cumming (2009), where such confidence intervals do not overlap, we can infer that the two estimates are significantly different at the 5 per cent level.[6] As can be seen, for English Literature and English Language there clearly are no significant differences between the estimated grade boundaries from different components. That is, while different components would indeed lead to different grade boundaries at whole qualification level, the size of these differences is no larger than we would expect given the quantifiable uncertainty in the CJ methods. That said, particularly for English Literature, some of the confidence intervals are very wide, which would restrict their practical usefulness operationally.

For Psychology, larger differences in the estimated boundaries are evident. For example, based on CJ evidence from paper 2, we would set the qualification A grade boundary at 186 marks. In contrast, based on paper 3 it would be set at 206 marks. These differences reflect the discrepancies already discussed earlier (Figure 2) in the evidence from different papers regarding the size of the difference in ability between the candidates that took the qualifications in different series.

The differences in grade boundaries (Figure 3, Table 4) are close to statistically significant and had we shown results at all grades rather than the key ones only, slightly larger, and statistically significant differences would have been visible. As such, we are confident in stating that it is possible for CJ evidence from different components to lead to significantly different results.

6 Similarly, Goldstein and Healy (1995) suggest that creating confidence intervals with estimates plus or minus 1.39 times the standard errors can ensure that the intervals for significantly different estimates will not overlap. This is equivalent to recommending the use of 84 per cent confidence intervals.

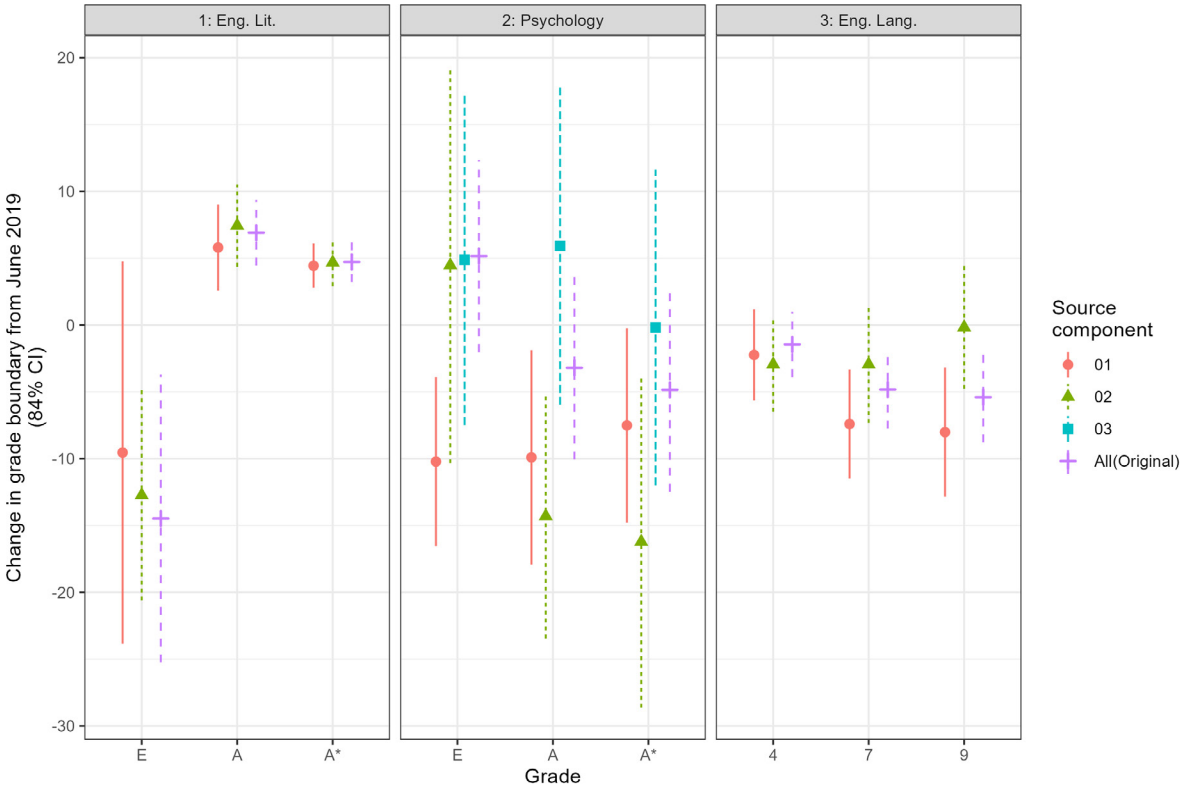


Figure 3: Changes in grade boundaries from June 2019 to autumn 2020 based on CJ evidence for each individual component alone and also based on all CJ evidence combined (with 84 per cent confidence intervals).

Aside from looking at the differences between estimated grade boundaries, it is also worth comparing the standard errors of grade boundaries from individual components to those based on the full set of CJ evidence. From Table 4 it can be seen that the confidence intervals of qualification grade boundaries derived from CJ evidence from a single component are mostly wider than those that combine all the CJ evidence. Specifically, the median width of confidence intervals from using all CJ evidence was plus or minus 4.5 marks. In contrast, the median width of confidence intervals using evidence from a single component was plus or minus 7 marks.

Conclusion

The aim of this research was to explore whether the use of CJ in awarding could be made more efficient by restricting judgements to a single component and then using the results to help infer grade boundaries for the qualification overall. Having compared the results of using CJ from a single component to using CJ exercises on all components, there are at least two reasons why we recommend that CJ in awarding should continue to incorporate separate exercises for each component:

- Relying on a single component leads to a noticeable decrease in precision. It should be noted that using CJ in awarding is already somewhat imprecise with analysis providing a range of scores that are consistent with judges' decisions rather than a single score.

- Relying on a single component effectively assumes that changes in the levels of candidate performance in one component are likely to be reflected in others. However, the analysis in this article has revealed that different components may suggest differing amounts of change in performance levels. This may be true even if our data indicates strong correlations between scores on different components. As such, if we accept that CJ results provide a realistic means of assessing changes in performance, we cannot assume that results from a single component are sufficient to infer how performance has changed on the qualification as a whole.

With regard to the issue of the loss of precision, it is possible that this might be addressed by increasing the size of the single CJ exercise, for example, by increasing the numbers of judges, scripts or comparisons included in the exercise. However, given the rate at which standard errors associated with CJ exercises fall with additional resources (see Benton, 2021) we would not expect this to provide a practical means to address this issue.

In terms of the evidence for using CJ in awarding at all, it would have been reassuring if we had found that the CJ results on every individual component in a qualification suggested the same level of change in performance among candidates. For example, this might have suggested that differences on all components were explicable in terms of a change in the general ability or prior attainment of the candidates entering a qualification. However, the fact that such consistency was not found for all three qualifications cannot be taken to imply a problem with the use of CJ for awarding. Particularly given the context in which the examinations were taken (a global pandemic), and the relatively small number of candidates entering autumn exams, it is genuinely plausible that changes in performance levels differ across components.

Further research might explore whether there are conditions where using a single component can be effective. Intuitively, we would expect the consistency of evidence from different papers to increase with greater overlap in the topics that they assess. Furthermore, it would be interesting to repeat the analysis in this article on data collected outside of the conditions of a global pandemic to see whether this leads to greater consistency of evidence. For example, we might speculate that the reason for the different patterns shown in Psychology paper 3 in our analysis is that the interruption to students' studies in 2020 meant they did not get to fully cover the topics in this paper.

Overall, this article provides little evidence as to whether the use of CJ in awarding is effective. Although this has been explored in various previous pieces of research (e.g., Benton et al., 2020; Benton et al., 2022; Curcin et al., 2019) it remains an open research question. However, this article does suggest that, if CJ is to be used in awarding activities, it is best if judges explicitly review performance on all of the different components. After all, we have seen that it is at least plausible that evidence from different components may lead to different results. Furthermore, relying on a single CJ exercise for a single component to grade an entire qualification decreases precision and makes additional assumptions that may or may not be correct.

Acknowledgement

With thanks to all those involved in the original data collection on which this research is based.

References

Benton, T. (2021). Comparative judgement for linking two existing scales. *Frontiers in Education, 6*, 775203.

Benton, T., Cunningham, E., Hughes, S., & Leech, T. (2020). *Comparing the simplified pairs method of standard maintaining to statistical equating*. Cambridge Assessment Research Report.

Benton, T., & Elliott, G. (2015). The reliability of setting grade boundaries using comparative judgement. *Research Papers in Education*, *31*(3), 352–376.

Benton, T., Gill. T., Hughes, S., & Leech. T. (2022). A summary of OCR's pilots of the use of Comparative Judgement in setting grade boundaries. *Research Matters: A Cambridge University Press & Assessment publication*, 33, 10–30.

Bramley, T. (2007). Paired comparison methods. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–294). Qualifications and Curriculum Authority.

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine, 28*(2), 205–220.

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots*. Ofqual report Ofqual/19/6575.

Goldstein, H., & Healy, M. J. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *158*(1), 175–177.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

Accessibility of GCSE science questions that ask students to create and augment visuals: Evidence from question omit rates

Santi Lestari (Research Division)

Introduction

Visual representations including graphs, diagrams, images and illustrations are prevalent in science texts and play a key role in science communication (Trumbo, 1999). They are often used to support verbal descriptions or explanation of complex scientific concepts and processes (Wang & Wei, 2024). Scientific visual literacy has therefore received considerable attention in science education and has been a feature in science education reform in several jurisdictions (LaDue et al., 2015; Wang & Wei, 2024). Scientific visual literacy encompasses not only the ability to interpret scientific visual representations but also to create them. There are ample arguments for, and evidence of, how visual representation construction is core to science learning (e.g., Ainsworth et al., 2011; Prain & Tytler, 2012; Tytler et al., 2018; Tytler et al., 2020). Therefore, including questions which require students to create visual representations in exams has been strongly advocated (Unsworth & Herrington, 2023; Wang & Wei, 2024). Chang et al. (2020) also note that drawing is a powerful method to assess students' understanding of scientific concepts and emphasise the advantages of requiring students to draw, rather than to write, for assessing certain concepts in science. As such, the Department for Education GCSE Science subject content document includes not only interpreting data presented in visual forms, but also communicating scientific observations and concepts through the creation of visual representations, as skills to be developed and assessed under "working scientifically" (Department for Education, 2015).

Exam question features could affect students' ability to engage with an exam question, i.e., to understand the question and subsequently to respond to it to demonstrate their knowledge, skills and understanding (Crisp & Macinska, 2020). For exam questions which require students to create a visual representation or augment a partially provided one, question features that could potentially influence students' performance include the layout of the visual representation and the amount of answer space. For example, if the answer space is too restricted due to certain layout formatting of the question, students might struggle to fit their answer within the space and therefore their ability to

demonstrate their understanding could be compromised. In short, design features of an exam or exam question could present accessibility issues which in turn may weaken the validity of test score inferences (Beddow, 2012).

Given the important influence of accessibility on validity, awarding organisations are required to ensure accessibility of their exams (Ofqual, 2022). OCR has a set of accessibility principles for GCSE Science to provide guidance in test construction and, thus, to ensure that all students can demonstrate their knowledge, skills and understanding (OCR, 2018a, 2018b). Some of these principles specifically relate to the use of visuals (i.e., the inclusion, placement and layout of visuals), and one principle in particular concerns questions which require students to do something with a visual (i.e., the visual will be centred with sufficient space around it to allow students to fit in their response). Such principles can help ensure that questions are as accessible as possible for candidates, but there is also a role for ongoing evaluation of the accessibility of exam questions.

There are multiple ways to investigate the accessibility of exams and exam questions. One method is by collecting expert judgements. For example, Beddow et al. (2013) asked test development experts to review exam questions using the *Accessibility Rating Matrix*. One of the elements assessed in the matrix is the use of visuals (e.g., the complexity of visuals and the placement of visuals). Another method involves gathering students' perspectives. For example, Crisp and Macinska (2020) interviewed students to gather their perspectives on the accessibility of GCSE Science questions. Other methods could involve conducting linguistic analyses of exam questions, as exemplified by Beauchamp and Constantinou (2020).

In this article, we argue that analyses of question omit rates could provide information about question accessibility. The omit rate for a question refers to the proportion of students who did not provide a response. Given that general qualifications in England use positive marking and, thus, there is no penalty for providing an incorrect answer, it is in the candidate's best interest to try to answer all questions (Sarac & Loken, 2023).

While research mostly focuses on the quality of student responses in an exam (i.e., correct, partially correct and incorrect responses), omit rates could also provide additional information about the exam (Papanastasiou, 2020) and could be useful to investigate various aspects of exams such as speededness (e.g., Walland, 2024) and differential test functioning (e.g., Ben-Shakhar & Sinai, 1991). Omit rates, however, have not been commonly used to investigate exam accessibility even though they could be an indicator of accessibility barriers. If certain questions or question types have systematically high omit rates, this could indicate potential access barriers. It could be argued that the nature and level of the demands of questions also contributes to variability in question omit rates. Referring to the CRAS scale of demands[1] (Pollitt et al., 2007), questions that require students to create a visual or augment a partially provided one can be considered to have a distinct and potentially higher level of strategy demand. In particular, response strategy demand, whereby students are required to organise how to

1 CRAS stands for Complexity, Resources, Abstractness and Strategy.

communicate their response through a visual representation, could be affected. Performance data (i.e., correct, partially correct and incorrect responses) may mask these potential accessibility barriers.

While the current research explores the use of omit rates as a possible indicator of accessibility, it is important to be aware that various factors could contribute to questions being omitted. Previous research has shown that omit rates can be influenced by characteristics of the student (e.g., ability level, gender, cultural background), characteristics of the exam and exam question (e.g., exam content, question format, question difficulty, question position) and interactions of the two. Examining the pattern of question omit rates in a low-stakes multiple-choice reading comprehension assessment, Clemens et al. (2015) found that students from lower performing subgroups had higher omit rates than those from higher performing subgroups, especially on the questions towards the end of the test.

In a larger-scale study involving high-stakes GCSE exams in biology, chemistry, physics, science and mathematics with mixed question formats, Walland (2024) found that omit rates for questions towards the end of exam papers were much higher for students from the lowest achieving subgroup than for those from the other subgroups. The foundation tier papers also had higher omit rates for questions towards the end of the papers than the higher tier papers. While skipping difficult questions could be an indicator of students' use of test-taking strategies, higher omit rates for questions towards the end of the test and especially for lower attaining students are more indicative of this particular group of students not being able to finish the test. This could be because lower attaining students might tend to take more time to attempt questions more generally, including those presented earlier in the test, as they find them harder than their higher attaining peers would. This would result in lower attaining students being more likely to leave questions towards the end of the test unanswered. In addition, it could also be the case that lower attaining students do not have sufficient knowledge, skills and understanding to be able to make an attempt at these later questions, given the tendency for a rough progression of question difficulty through a paper. Therefore, the patterns of question omit rates identified in Clemens et al. (2015) and Walland (2024) seem more likely to be due to the interaction between students' ability level and the difficulty of the subject content in the questions than due to accessibility issues.

Students' gender has also been found to interact with omit rates. Male students were typically found to omit fewer questions than female students in multiple-choice tests (e.g., Ben-Shakhar & Sinai, 1991). However, these patterns of omit rates across genders could vary across question formats and subjects. Matters and Burnett (1999), researching the high-stakes Queensland Core Skills Test, found that for multiple-choice questions omit rates in general were very small and the difference across gender categories was negligible. For constructed response questions, however, omit rates were higher and male students omitted more questions than female students. In von Schrader and Ansley's (2006) analysis of the high-stakes Iowa Tests of Basic Skills and Iowa Tests of Educational Development, female students tended to omit more questions in the mathematics exam while male students tended to omit more questions in the reading and vocabulary exams.

The research reported in the current article examined the accessibility of GCSE Science questions involving the creation and augmentation of visuals (e.g., adding an element to a partially provided diagram) by analysing the patterns of question omit rates.

The main question guiding the research was:

- Is there empirical evidence of atypically high omit rates for GCSE Science items that require diagram creation or augmentation, which could indicate a potential accessibility issue?

To address the research question, omit rates for questions involving creating or augmenting visuals were compared to those for questions without, and these comparisons were conducted across tiers (i.e., foundation and higher tiers), subjects (i.e., biology, physics, chemistry and combined science), question position within a paper, maximum marks and facility values, as well as by gender and attainment group. Comparisons across different question attributes and candidate characteristics were made to help differentiate factors other than question accessibility that may have contributed to omit rates.

Method

Data

In this research, we used the item-level data (marks or omission information for each candidate on each item) from eight OCR GCSE Science specifications[2] from the June 2023 exam series. Each specification had different numbers of papers, and in total there were 44 papers.[3]

Approach to item categorisation

While visual representations are often used in GCSE Science exams, in this study we specifically focused on questions which require students to create a visual representation or augment a partially provided one (e.g., drawing a line of best fit on a graph or completing a diagram). For the sake of brevity, such questions are referred to as "items with diagram(s)" in the remainder of this article.

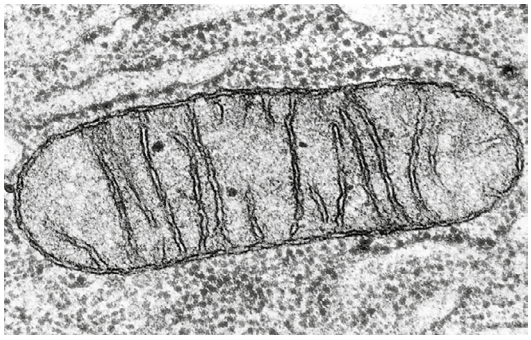
To illustrate, Item b(i) in Figure 1 is an example of an item with a diagram because it requires students to create a scientific drawing. Items 19a(i) and (ii) in Figure 2 are also both examples of items with diagrams because they require students to augment a partially provided graph. Conversely, although Item b in Figure 3 is based on a diagram, it is not considered an item with a diagram in this

2 The eight specifications were: Science A, Combined (9–1) – Gateway Science Suite J250; Biology A (9–1) – Gateway Science Suite J247; Chemistry A (9–1) – Gateway Science Suite J248; Physics A (9–1) – Gateway Science Suite J249; Science B, Combined (9–1) – Twenty First Century Science Suite J260; Biology B (9–1) – Twenty First Century Science Suite J257; Chemistry B (9–1) – Twenty First Century Science Suite J258; Physics B (9–1) – Twenty First Century Science Suite J259.

3 The data used in this research was collected as part of the usual marking and processing of candidates' examination scripts. Data has been stored and used in line with Cambridge University Press & Assessment's Data Privacy notice (https://www.cambridge.org/legal/candidate-privacy-notice).

study because it requires students to explain a process rather than creating or augmenting a diagram.

(b) The image is of a mitochondrion.



(i) Draw the mitochondrion in the box. Your drawing should be a scientific drawing.

[2]

Figure 1: A sample item requiring students to create a visual representation (categorised as "item with diagram(s)")[4]

4 Source: https://www.ocr.org.uk/Images/704945-question-paper-paper-1.pdf

19 A student investigates the effect of pH on an enzyme called catalase.
Catalase breaks down hydrogen peroxide into water and oxygen.

The student collects the oxygen produced by the reaction.
The table shows their results.

pH	Volume of oxygen collected (cm^3)
2	1
4	12
6	24
8	26
10	8

(a) (i) Plot a graph of the results. **[2]**

(ii) Draw a line of best fit. **[1]**

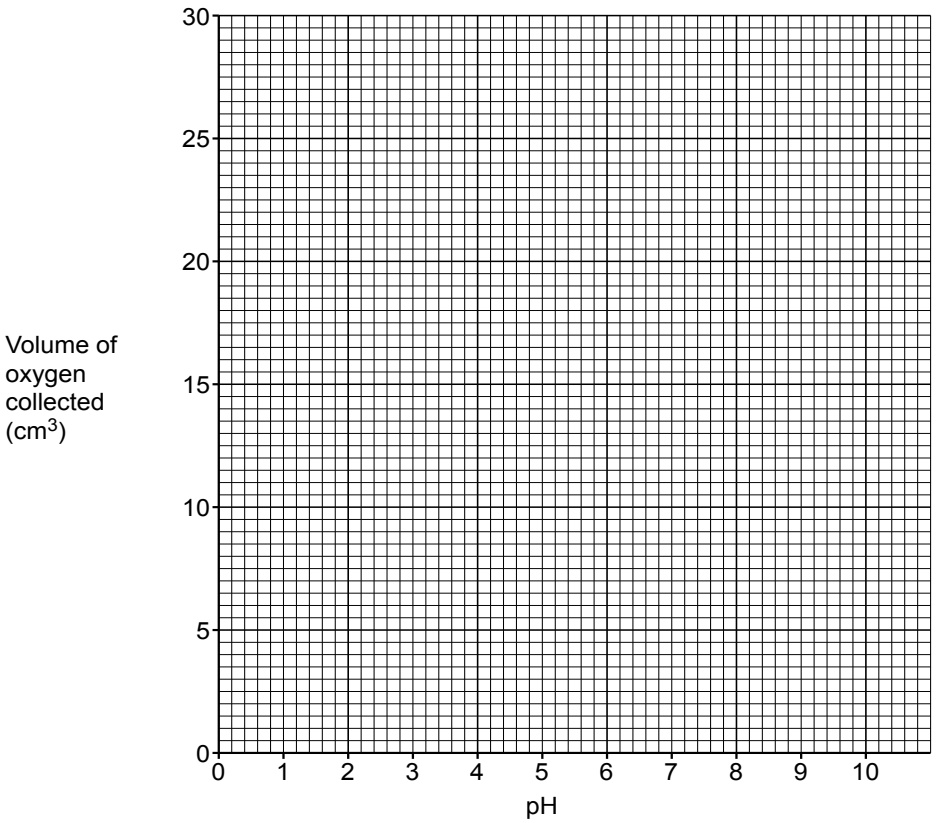
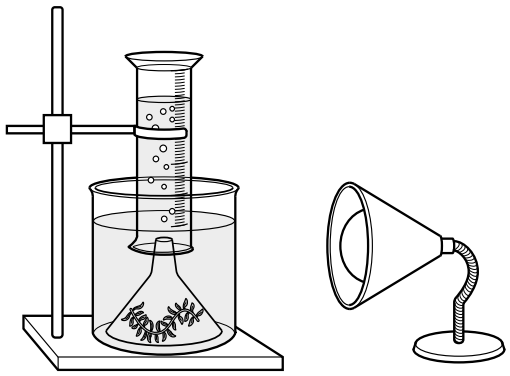


Figure 2: Sample items requiring students to augment a partially provided visual representation (both items categorised as "item with diagram(s)")[5]

5 Source: https://www.ocr.org.uk/Images/704945-question-paper-paper-1.pdf

(b) The student sets up a second experiment using the equipment in **Fig. 3.2**.

Fig. 3.2



Explain how this second experiment will improve the quality of the data collected to measure the rate of photosynthesis.

.....

.....

.....

..... **[2]**

Figure 3: A sample item requiring students to explain a scientific process illustrated in a diagram (categorised as "item without diagrams")[6]

Procedure

The item-level data from the exam papers were processed in three data preparation steps:

1. **Coding of items.**
   As described and exemplified in the previous section, items were binary coded as **"item with diagram(s)"** or **"item without diagrams"**. It should be noted that each paper had only small numbers of items with diagrams (typically three to four), and of 44 papers, only six had more than five items with diagrams.

2. **Removal of multiple-choice item data.**
   Initial exploration of the data suggested that multiple-choice items tended to have zero or close to zero omit rates, which is unsurprising considering the possibility of guessing and the question position (first section of the paper). These very low omit rates would skew the omit rate distribution for a whole paper. Therefore, the multiple-choice item data were excluded from the analysis.[7]

6 Source: https://www.ocr.org.uk/Images/705023-question-paper-combined-science.pdf
7 Only J247, J248, J249 and J250 specifications had multiple-choice items. There were 10 multiple-choice items in each paper in J250, and 15 multiple-choice items in each paper in J247, J248, and J249.

3. **Calculations based on item-level data.**
   Firstly, the **omit rate** needed to be calculated for each item in each paper. Omit rate refers to the proportion of students who did not attempt an item. The value ranges from 0 to 1, with 0 indicating that all students attempted the item and 1 indicating that no student attempted the item.

   To further examine the patterns of omit rates for different gender and attainment groups, disaggregated omit rates also had to be calculated. Item omit rates for each gender group were calculated.[8] Students were also classified into attainment quartiles based on the total marks they achieved in each paper. Then, the item omit rate for each attainment quartile was calculated.

   As the papers had different numbers of items and maximum marks, **item position** within a paper needed to be standardised. Item position was therefore defined as the proportion of how far through the paper an item was in terms of the paper maximum mark. The value ranged from 0 to 1 and was classified into quintiles (i.e., 0.00–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80 and 0.81–1.00).

   Item maximum mark in the papers included in this study ranged from 1 to 6. However, there were very few items with a maximum mark of 4 to 6. Therefore, maximum marks of 4 to 6 were grouped together into "4 or above".

   The **facility value** for each item in each paper also needed to be calculated. Facility value refers to the mean mark on the item as a proportion of maximum mark and is a useful measure of item difficulty on exams where all the items are compulsory. The value ranges from 0 to 1, with 0 indicating that no marks were scored by any students and 1 indicating that all students achieved maximum marks on the item. To facilitate analysis, item facility value was classified into quintiles (i.e., 0.00–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80 and 0.81–1.00).

   Finally, all of the variables calculated based on the item-level data for all the 44 GCSE Science papers were combined into a single dataset for further analyses.

The dataset was analysed using descriptive statistics to address the research question. More specifically, the omit rates for items with diagrams were compared with those for items without diagrams. These comparisons were conducted across:

a. tiers (foundation and higher tiers)
b. subjects (i.e., biology, physics, chemistry and combined science)
c. item positions within a paper
d. item maximum marks
e. item facility values (a measure of item difficulty level)

8 The analysis across gender groups did not include data for candidates with no gender information due to the extremely small size of this group.

We also examined the patterns across students from different gender and attainment groups. Boxplots were used to visualise the results. Both the descriptive statistics analyses and boxplot generation were conducted in RStudio (Posit team, 2023).

Results

The results are presented for each of the five aspects (i.e., omit rates by tier, subject, item position, item maximum mark and item facility value) in turn.

Boxplots are used to visualise the distribution of item omit rates and accompanied by tables showing their associated descriptive statistics. In a boxplot, the horizontal line dividing the box into two represents the median value. The line on the lower edge of the box represents the lower quartile, and the line on the higher edge represents the upper quartile. The lines extending from the box, known as the whiskers, represent the variability in the dataset beyond the lower and upper quartiles. The individual dots represent the outliers.

It is important to remember that the proportion of items with diagrams in each paper was generally very small (8 per cent on average), so the results should be interpreted cautiously. It is recommended to consult the descriptive statistics tables that contain the number of items for each category.

Omit rate by tier

As shown in Figure 4 and Table 1, omit rates were generally higher in the foundation tier papers than in the higher tier papers, for both items with and without diagrams. In fact, omit rates for items in the higher tier papers were all very low on average, making it difficult to examine differences in omit rates across tiers as well as across item types within the higher tier papers. Within the foundation tier, although the median omit rate for items with diagrams appeared slightly lower than that for items without diagrams, this difference was still too small to be meaningful.

Analysis of the disaggregated data by attainment group (quartiles) based on candidate overall performance in each paper showed that for both the foundation and higher tiers, omit rates were higher in the lower attainment groups and decreased in the higher attainment groups (see Figure 5 and Table 2). Omit rates were considerably higher for the lowest attainment group (Q1) in the foundation tier than the other quartiles. This was true for both items with and without diagrams, with no meaningful differences observed. Omit rates in the foundation tier papers for candidates in Q2, Q3 and Q4 were broadly comparable to omit rates in the higher tier papers.

Further analysis of the disaggregated data by gender showed that male candidates in the foundation tier papers generally had a higher propensity to omit items than their female peers did, and this was true for both items with and without diagrams, although the difference might be negligible (see Figure 6 and Table 3).

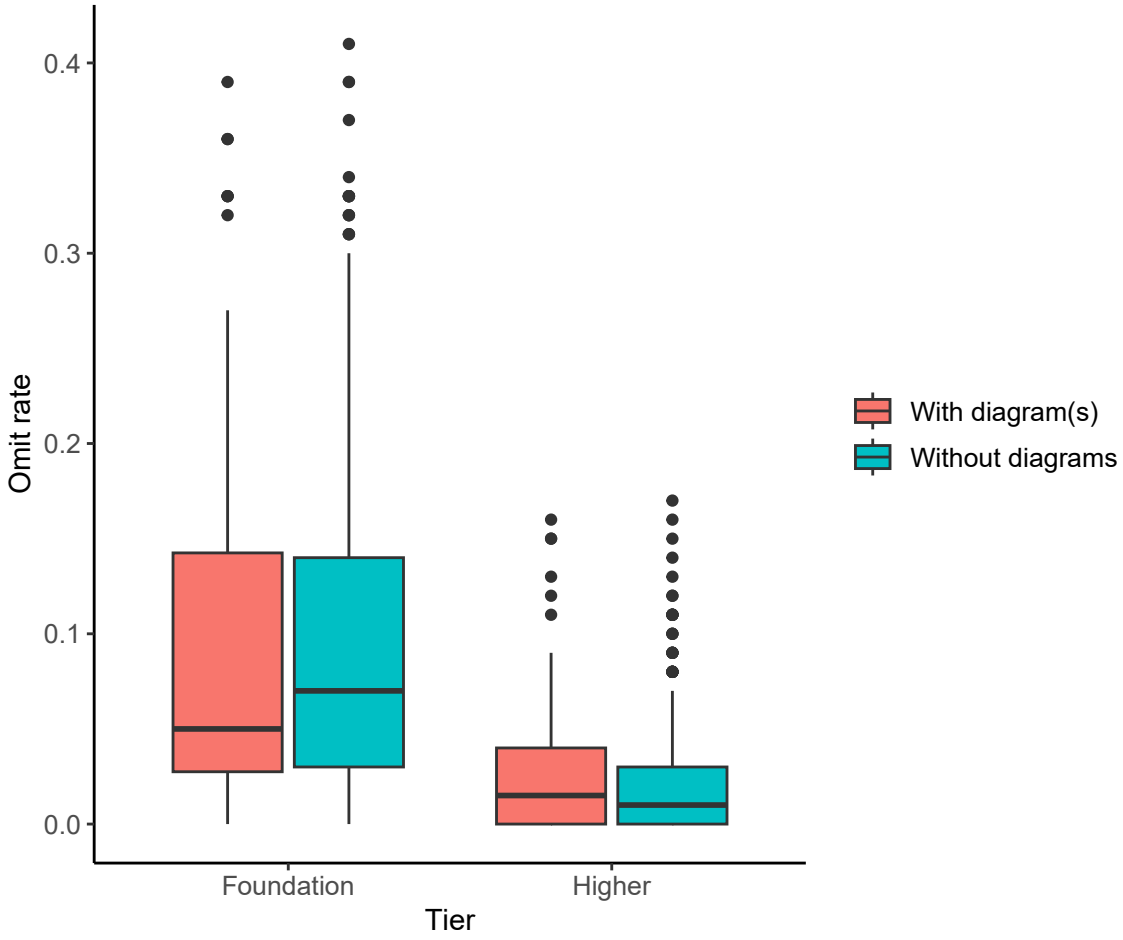


Figure 4: Omit rate by tier

Table 1: Omit rate descriptive statistics, by tier

Tier	Item type	Number of items	Min	Max	Median	Mean	SD
Foundation	With diagram(s)	84	0	0.39	0.05	0.10	0.10
	Without diagrams	793	0	0.41	0.07	0.09	0.08
Higher	With diagram(s)	76	0	0.16	0.01	0.03	0.04
	Without diagrams	718	0	0.17	0.01	0.02	0.02

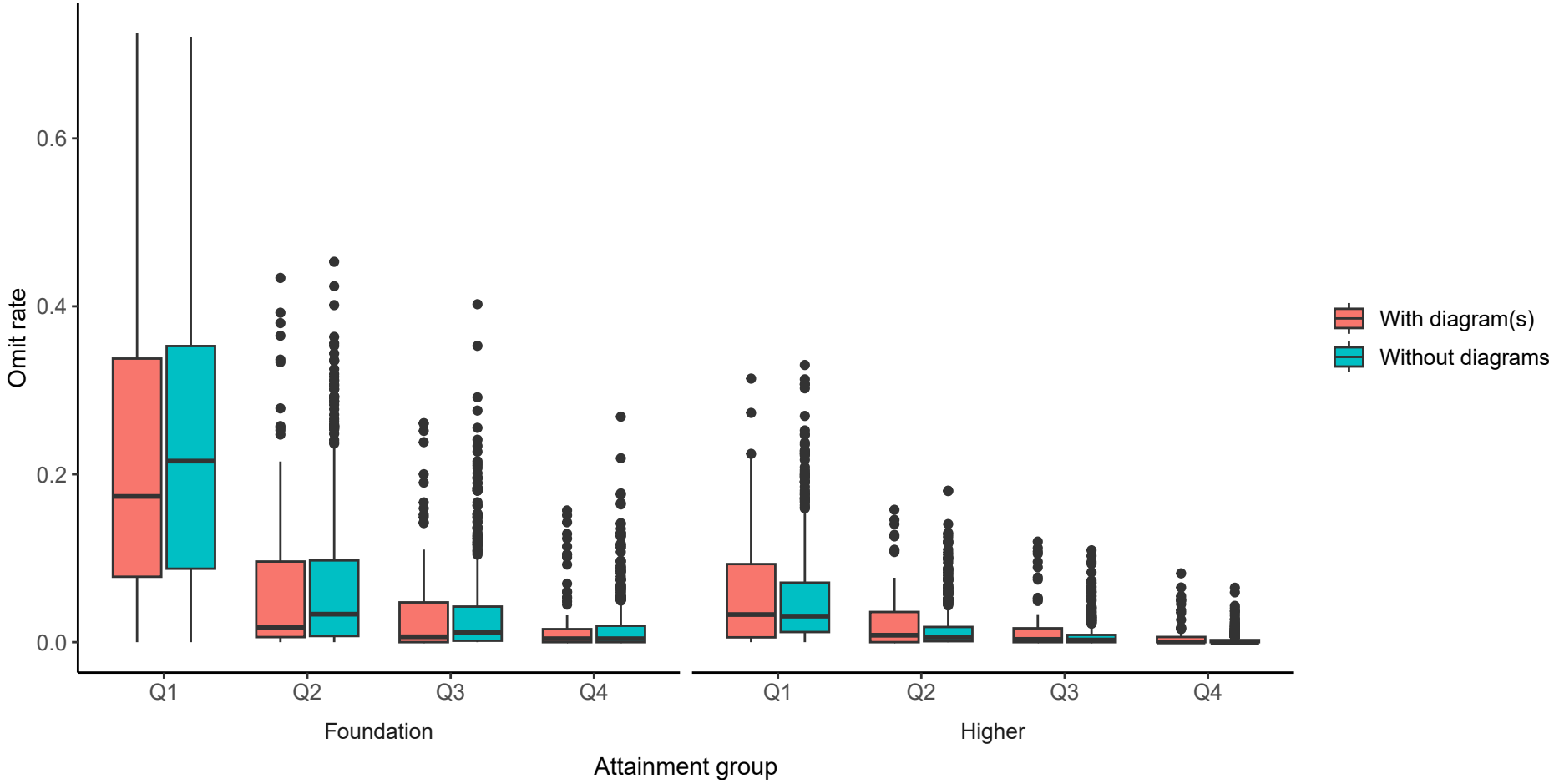


Figure 5: Omit rate by tier and attainment group (Q1 being the lowest attaining quartile and Q4 the highest attaining quartile)

Table 2: Omit rate descriptive statistics, by tier and attainment group (Q1 being the lowest attaining quartile and Q4 the highest attaining quartile)

Tier	Attainment group	Item type	Number of items	Min	Max	Median	Mean	SD
Foundation	Q1	With diagram(s)	84	0	0.73	0.17	0.24	0.20
		Without diagrams	793	0	0.72	0.22	0.24	0.17
	Q2	With diagram(s)	84	0	0.43	0.02	0.08	0.11
		Without diagrams	793	0	0.45	0.03	0.07	0.08
	Q3	With diagram(s)	84	0	0.26	0.01	0.04	0.07
		Without diagrams	793	0	0.40	0.01	0.03	0.05
	Q4	With diagram(s)	84	0	0.16	0	0.02	0.04
		Without diagrams	793	0	0.27	0	0.02	0.03
Higher	Q1	With diagram(s)	76	0	0.31	0.03	0.06	0.07
		Without diagrams	718	0	0.33	0.03	0.05	0.06
	Q2	With diagram(s)	76	0	0.16	0.01	0.03	0.04
		Without diagrams	718	0	0.18	0.01	0.02	0.02
	Q3	With diagram(s)	76	0	0.12	0	0.02	0.03
		Without diagrams	718	0	0.11	0	0.01	0.01
	Q4	With diagram(s)	76	0	0.08	0	0.01	0.02
		Without diagrams	718	0	0.06	0	0	0.01

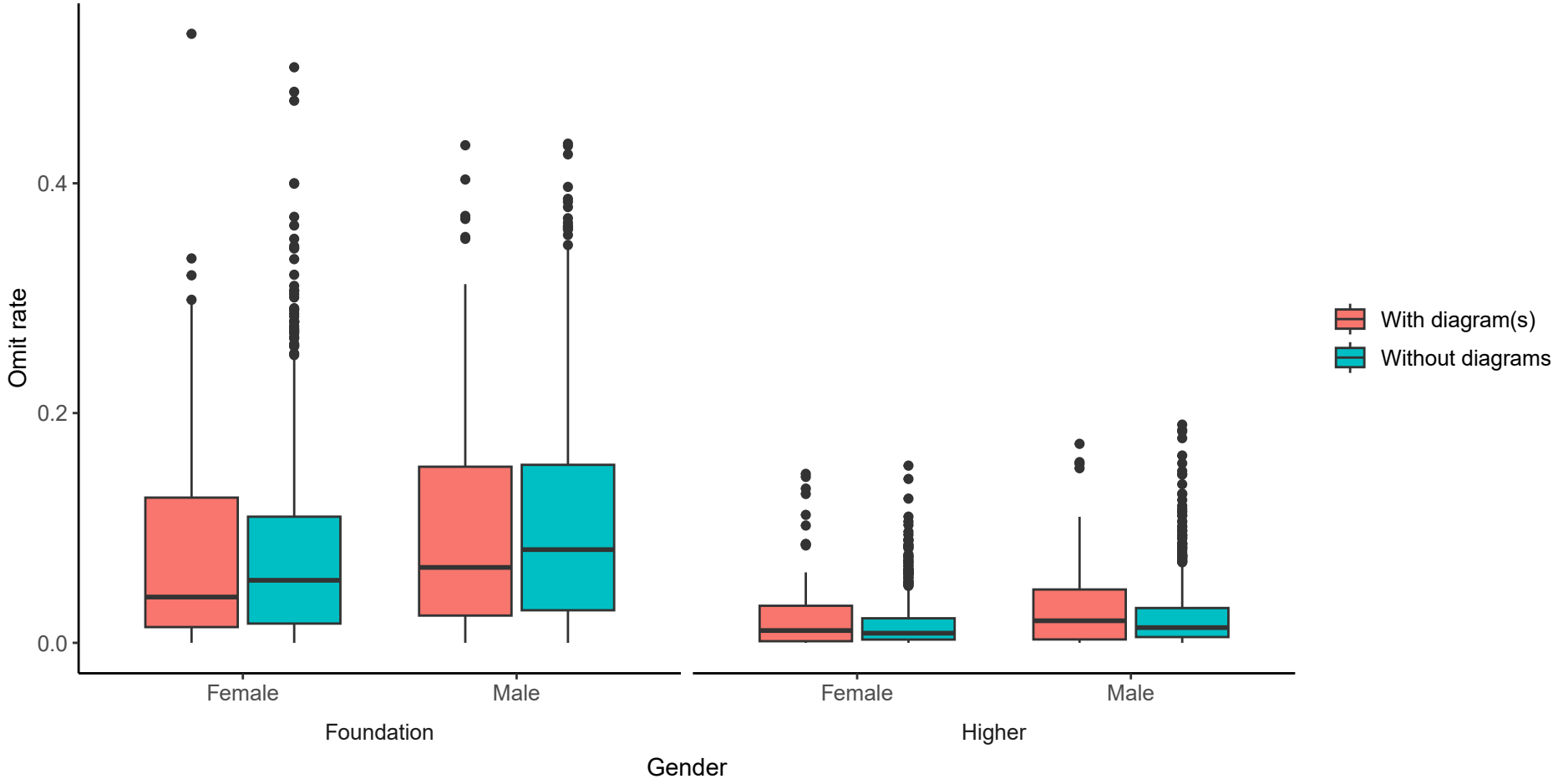


Figure 6: Omit rate by tier and gender

Table 3: Omit rate descriptive statistics, by tier and gender

Tier	Gender	Item type	Number of items	Min	Max	Median	Mean	SD
Foundation	Female	With diagram(s)	84	0	0.53	0.04	0.09	0.10
		Without diagrams	793	0	0.50	0.05	0.08	0.08
	Male	With diagram(s)	84	0	0.43	0.07	0.11	0.11
		Without diagrams	793	0	0.43	0.08	0.10	0.09
Higher	Female	With diagram(s)	76	0	0.15	0.01	0.03	0.04
		Without diagrams	718	0	0.15	0.01	0.02	0.02
	Male	With diagram(s)	76	0	0.17	0.02	0.03	0.04
		Without diagrams	718	0	0.19	0.01	0.02	0.03

Omit rate by subject

Figure 7 and Table 4 show the omit rates by tier and subject. Omit rates varied across papers assessing different subjects. Note that there was only one combined science paper in each tier, hence the small numbers of items in these papers. Chemistry papers overall had higher omit rates than other papers, particularly in the foundation tier. Furthermore, chemistry items with diagrams in the foundation tier appeared to have higher omit rates than those without diagrams. Conversely, biology items with diagrams in the foundation tier tended to have lower omit rates than those without diagrams. There were no substantial differences in omit rates for items with and without diagrams in physics papers. While Figure 7 shows differences in the median omit rates between items with diagrams and items without diagrams in the combined science papers, these differences were not meaningful due to the low numbers of items with diagrams (i.e., only five items in the foundation tier and three items in the higher tier).

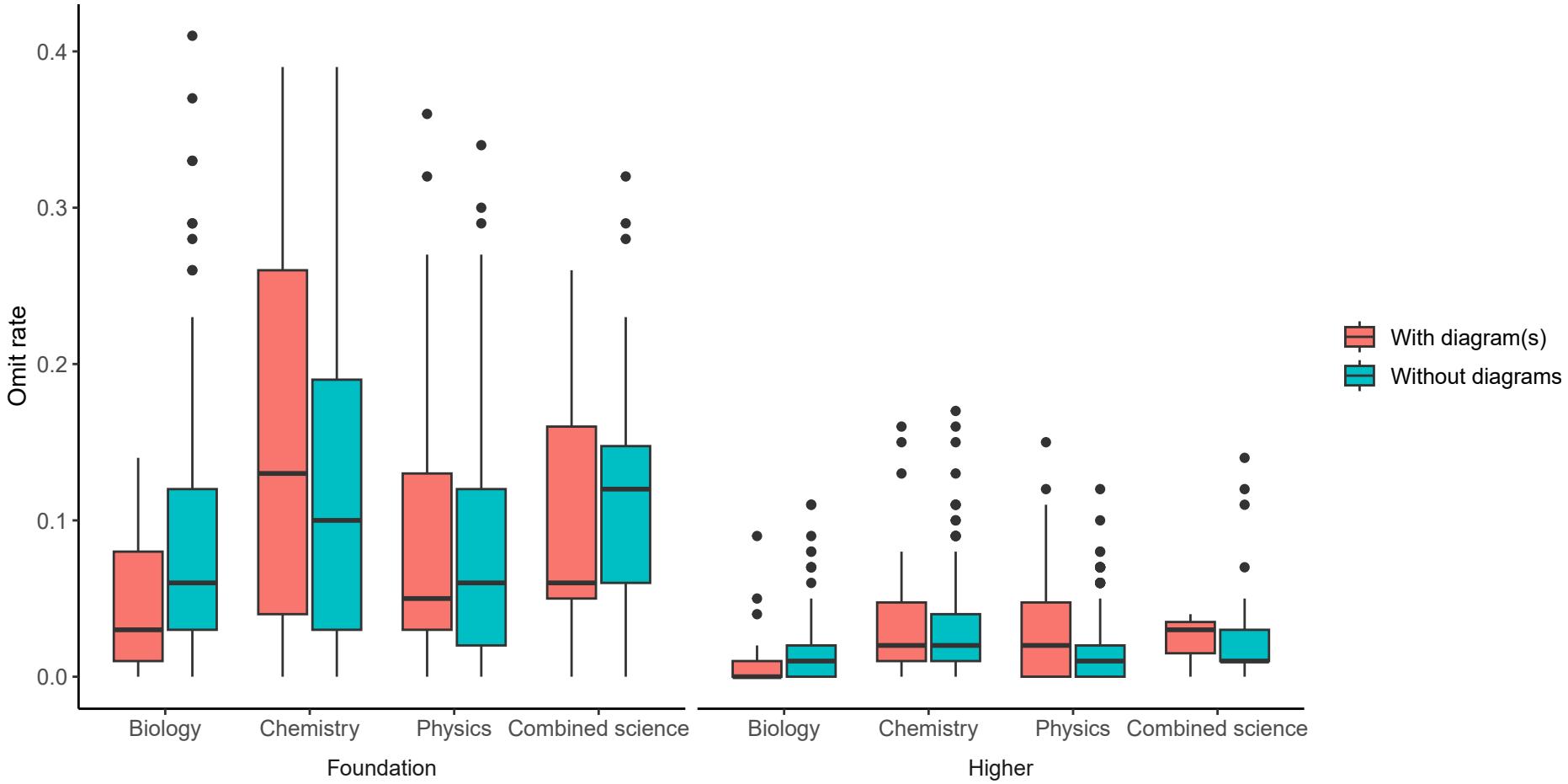


Figure 7: Omit rate by subject

Table 4: Omit rate descriptive statistics, by subject

Tier	Subject	Item type	Number of items	Min	Max	Median	Mean	SD
Foundation	Biology	With diagram(s)	21	0	0.14	0.03	0.04	0.04
		Without diagrams	243	0	0.41	0.06	0.08	0.07
	Chemistry	With diagram(s)	21	0	0.39	0.13	0.16	0.13
		Without diagrams	185	0	0.39	0.10	0.12	0.10
	Physics	With diagram(s)	37	0	0.36	0.05	0.09	0.09
		Without diagrams	333	0	0.34	0.06	0.08	0.07
	Combined science	With diagram(s)	5	0	0.26	0.06	0.11	0.10
		Without diagrams	32	0	0.32	0.12	0.12	0.08
Higher	Biology	With diagram(s)	21	0	0.09	0	0.01	0.02
		Without diagrams	234	0	0.11	0.01	0.01	0.02
	Chemistry	With diagram(s)	22	0	0.16	0.02	0.04	0.05
		Without diagrams	172	0	0.17	0.02	0.03	0.03
	Physics	With diagram(s)	30	0	0.15	0.02	0.03	0.04
		Without diagrams	285	0	0.12	0.01	0.02	0.02
	Combined science	With diagram(s)	3	0	0.04	0.03	0.02	0.02
		Without diagrams	27	0	0.14	0.01	0.03	0.04

Omit rate by item position

Figure 8 and Table 5 show omit rates by tier and item position within a paper. There appeared to be an overall tendency that omit rates increased as the paper progressed. In other words, items towards the end of the paper tended to have slightly higher omit rates than those preceding them. For foundation tier papers, in particular, there was more variability in omit rates for items towards the end of the paper. From the middle to the end of the foundation tier paper, items with diagrams appeared to have lower omit rates than items without diagrams. However, it is hard to be sure that this trend was not random given the low numbers of items with diagrams.

Further analysis of omit rates by item position across different attainment groups (not reported in full here for reasons of brevity) indicated that the rate of omission for items towards the end of the paper was substantially higher for candidates in the lower attaining groups and especially in the foundation tier.

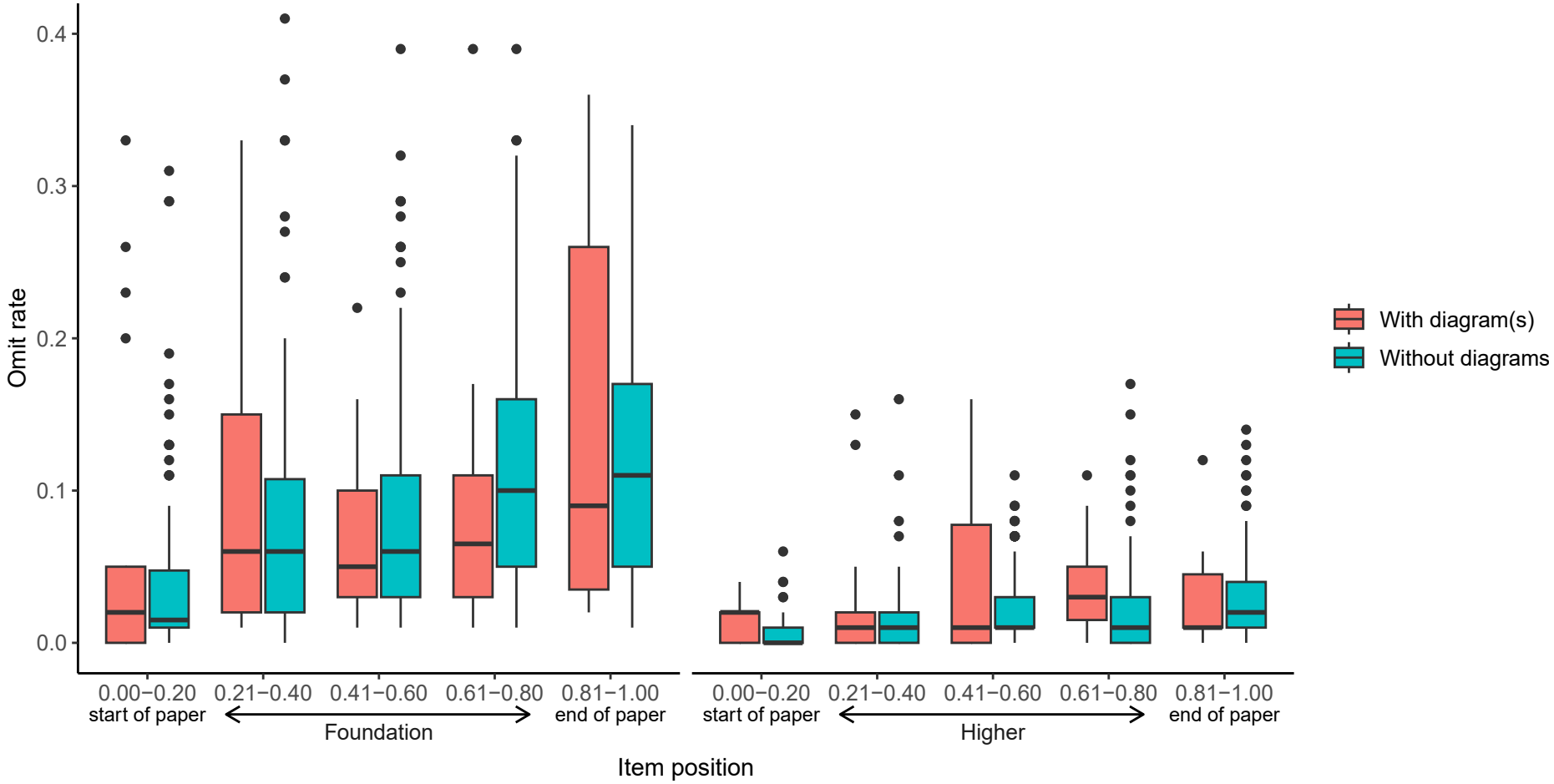


Figure 8: Omit rate by item position

Table 5: Omit rate descriptives, by tier and item position

Tier	Item position	Item type	Number of items	Min	Max	Median	Mean	SD
Foundation	0.00-0.20	With diagram(s)	17	0	0.33	0.02	0.07	0.11
		Without diagrams	102	0	0.31	0.01	0.04	0.06
	0.21-0.40	With diagram(s)	17	0.01	0.33	0.06	0.09	0.09
		Without diagrams	174	0	0.41	0.06	0.08	0.07
	0.41-0.60	With diagram(s)	19	0.01	0.22	0.05	0.07	0.06
		Without diagrams	171	0.01	0.39	0.06	0.08	0.07
	0.61-0.80	With diagram(s)	8	0.01	0.39	0.06	0.11	0.12
		Without diagrams	172	0.01	0.39	0.10	0.11	0.08
	0.81-1.00	With diagram(s)	23	0.02	0.36	0.09	0.14	0.12
		Without diagrams	174	0.01	0.34	0.11	0.12	0.07
Higher	0.00-0.20	With diagram(s)	9	0	0.04	0.02	0.02	0.02
		Without diagrams	104	0	0.06	0	0.01	0.01
	0.21-0.40	With diagram(s)	23	0	0.15	0.01	0.02	0.04
		Without diagrams	168	0	0.16	0.01	0.01	0.02
	0.41-0.60	With diagram(s)	14	0	0.16	0.01	0.04	0.06
		Without diagrams	154	0	0.11	0.01	0.02	0.02
	0.61-0.80	With diagram(s)	15	0	0.11	0.03	0.04	0.03
		Without diagrams	136	0	0.17	0.01	0.02	0.03
	0.81-1.00	With diagram(s)	15	0	0.12	0.01	0.03	0.03
		Without diagrams	156	0	0.14	0.02	0.03	0.03

Omit rate by item maximum mark

Figure 9 and Table 6 show omit rates by tier and item maximum mark. In the foundation tier, it is evident that items with a maximum mark of 1 that had diagrams tended to have higher omit rates than those items with the same maximum mark but without diagrams. Conversely, of items with a maximum mark of 3 and 4 or above, those that had diagrams tended to have lower omit rates than those that did not have diagrams.

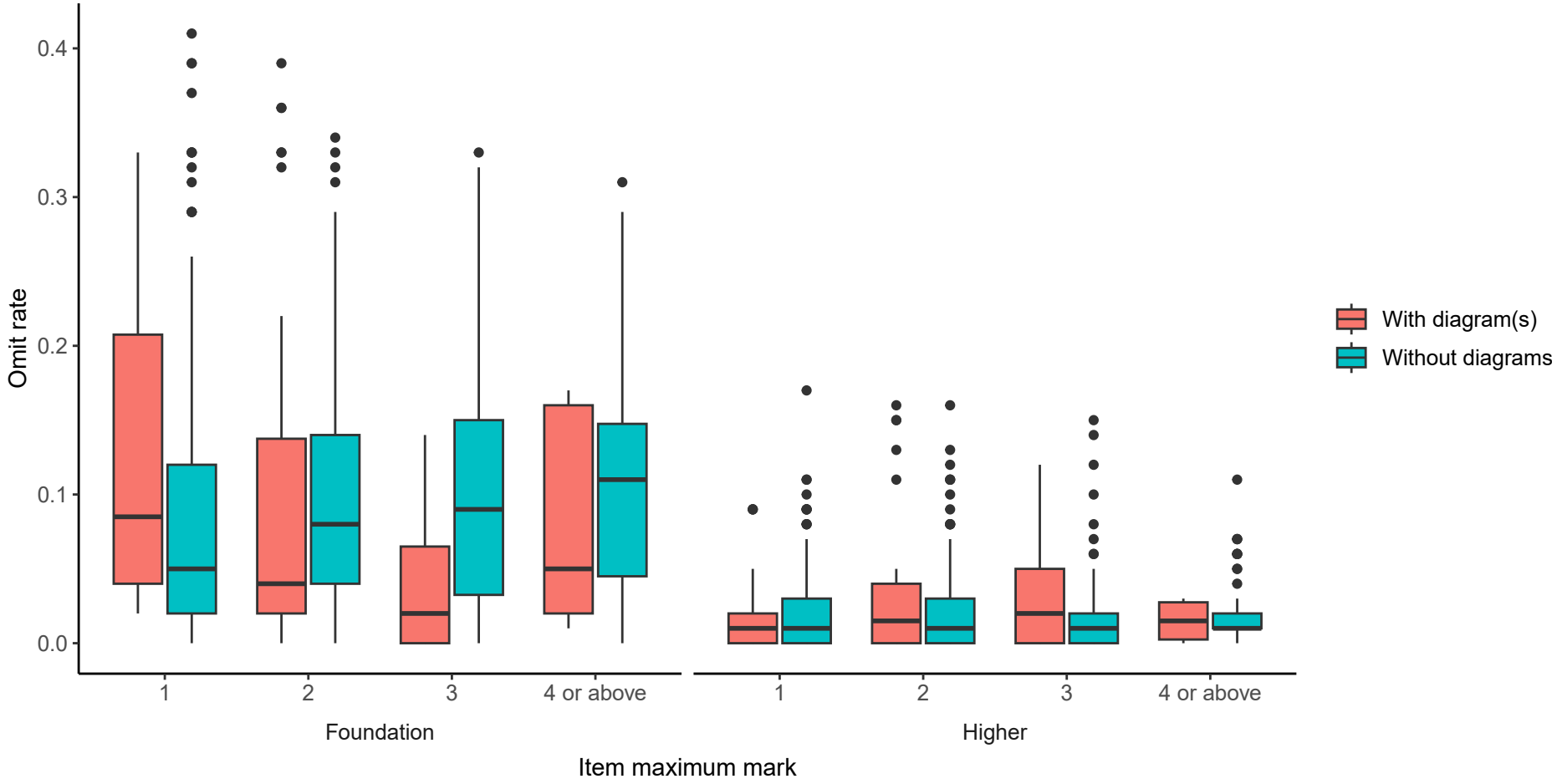


Figure 9: Omit rate by item maximum mark

Table 6: Omit rate descriptive statistics, by tier and maximum mark

Tier	Item max mark	Item type	Number of items	Min	Max	Median	Mean	SD
Foundation	1	With diagram(s)	28	0.02	0.33	0.08	0.12	0.10
		Without diagrams	347	0	0.41	0.05	0.08	0.08
	2	With diagram(s)	40	0	0.39	0.04	0.10	0.12
		Without diagrams	290	0	0.34	0.08	0.10	0.08
	3	With diagram(s)	11	0	0.14	0.02	0.04	0.05
		Without diagrams	110	0	0.33	0.09	0.10	0.08
	4 or above	With diagram(s)	5	0.01	0.17	0.05	0.08	0.08
		Without diagrams	46	0	0.31	0.11	0.11	0.08
Higher	1	With diagram(s)	21	0	0.09	0.01	0.02	0.03
		Without diagrams	251	0	0.17	0.01	0.02	0.02
	2	With diagram(s)	36	0	0.16	0.01	0.03	0.05
		Without diagrams	270	0	0.16	0.01	0.02	0.03
	3	With diagram(s)	13	0	0.12	0.02	0.03	0.04
		Without diagrams	134	0	0.15	0.01	0.02	0.02
	4 or above	With diagram(s)	6	0	0.03	0.01	0.01	0.01
		Without diagrams	63	0	0.11	0.01	0.02	0.02

Omit rate by item facility value

Figure 10 and Table 7 show the comparisons of omit rates between items with diagrams and items without diagrams at similar difficulty levels. Overall omit rates were higher for more difficult items (lower facility values) than for easier items. This is expected because the calculation of facility value also takes into account omissions. The omit rate of an item effectively limits the maximum possible facility value of an item since candidates who did not answer will have received 0 marks on the item (e.g., an item with an omit rate of 0.20, cannot have a facility value of more than 0.80).

It also appeared that items with diagrams tended to have higher omit rates than those at similar difficulty levels but without diagrams. This observation was particularly prominent for items with very low facility values (0.00–0.20), i.e., for very difficult items. While this could be taken to suggest that items with diagrams might have introduced access barriers, it is important to be cautious at interpreting this finding given that there were far fewer items with diagrams than without diagrams at this difficulty level — nine versus 176 in the foundation tier. A closer look showed that all of the nine items with diagrams were located towards the end of the papers. Given that items towards the end of the papers tended to have higher omit rates, it could be that the considerable difference in omit rates between items with diagrams and without was not solely or primarily due to the difficulty level and accessibility, but due to students not being able to reach these items.

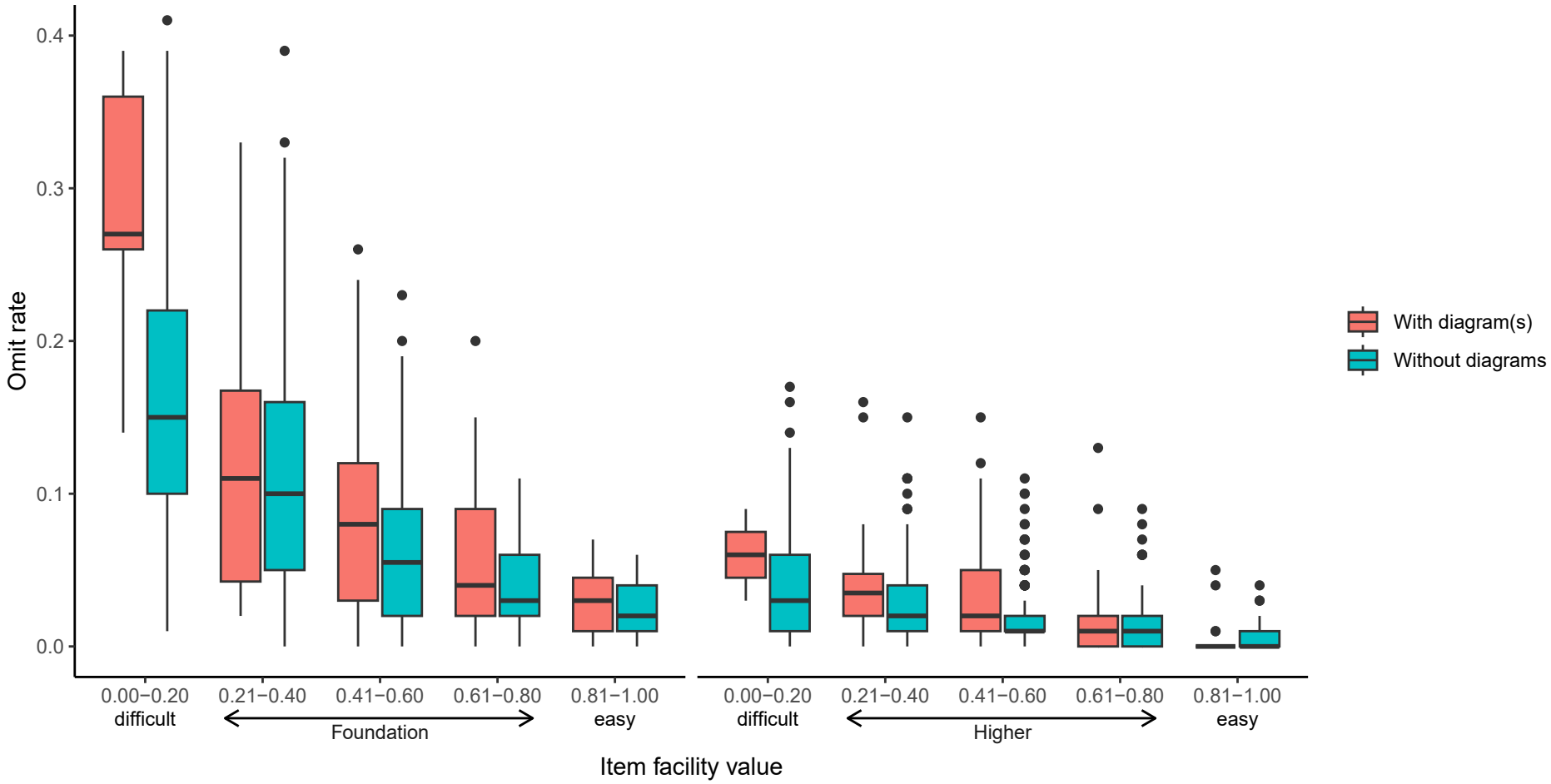


Figure 10: Omit rate by item facility value

Table 7: Omit rate descriptive statistics, by tier and facility value

Tier	Facility value	Item type	Number of items	Min	Max	Median	Mean	SD
Foundation	0.00-0.20	With diagram(s)	9	0.14	0.39	0.27	0.28	0.09
		Without diagrams	176	0.01	0.41	0.15	0.16	0.09
	0.21-0.40	With diagram(s)	14	0.02	0.33	0.11	0.14	0.12
		Without diagrams	217	0	0.39	0.10	0.11	0.07
	0.41-0.60	With diagram(s)	25	0	0.26	0.08	0.09	0.08
		Without diagrams	208	0	0.23	0.06	0.06	0.05
	0.61-0.80	With diagram(s)	17	0	0.20	0.04	0.06	0.05
		Without diagrams	139	0	0.11	0.03	0.04	0.03
	0.81-1.00	With diagram(s)	19	0	0.07	0.03	0.03	0.02
		Without diagrams	53	0	0.06	0.02	0.02	0.02
Higher	0.00-0.20	With diagram(s)	2	0.03	0.09	0.06	0.06	0.04
		Without diagrams	67	0	0.17	0.03	0.04	0.04
	0.21-0.40	With diagram(s)	18	0	0.16	0.04	0.04	0.04
		Without diagrams	164	0	0.15	0.02	0.03	0.03
	0.41-0.60	With diagram(s)	20	0	0.15	0.02	0.04	0.04
		Without diagrams	183	0	0.11	0.01	0.02	0.02
	0.61-0.80	With diagram(s)	17	0	0.13	0.01	0.02	0.04
		Without diagrams	197	0	0.09	0.01	0.01	0.02
	0.81-1.00	With diagram(s)	19	0	0.05	0	0.01	0.01
		Without diagrams	107	0	0.04	0	0	0.01

Discussion and conclusion

This research examined the accessibility of GCSE Science items that require students to create a visual or augment a partially provided one through analysing patterns of item omit rates.

Analyses of omit rates for items with and without diagrams by tier, subject, item position, item maximum mark and item facility value have shown that there was very little to no evidence that average omit rates were higher for items with diagrams compared to those without diagrams. Therefore, this research found no indication that items with diagrams in GCSE Science had potential accessibility issues.

Regardless of the item type (i.e., with or without diagrams), omit rates were overall higher for the foundation tier papers than for the higher tier papers. Furthermore, analysis of the disaggregated data by attainment group showed that omit rates were higher in the lower attainment groups and decreased in the higher attainment groups. In terms of patterns of omit rates by item position within a paper, there appeared to be a trend of increasing omit rates as a paper progressed; omit rates tended to be higher for items towards the end of the paper. These findings on omit rates by tier and attainment group, and omit rates by item position, taken together, support results from previous studies (Clemens et al., 2015; Walland, 2024), which suggested that student ability plays a role in the rate of item omission. Lower attaining students tended to omit more items and particularly items towards the end of the paper, most likely because they ran out of time to attempt these items. While this finding provides valuable insights into the likely cause of item omission, this does not indicate accessibility issues.

In terms of omit rates by subject, chemistry items with diagrams in the foundation tier papers were found to have slightly higher omit rates than those without diagrams. On the contrary, biology items with diagrams tended to have lower omit rates than those without. This finding provides an indication that subject area could also contribute to variability in omit rates in addition to or instead of item type (i.e., with or without diagrams). There could be various reasons for this, such as intrinsic differences in the kinds of visuals that learners are asked to create or augment in different subjects.

Items with diagrams that had a maximum mark of 1 in the foundation tier papers were found to have higher omit rates than those without, while items with diagrams that had a maximum mark of 3 and 4 or above had lower omit rates than those without. It could be speculated that candidates were more likely to attempt items with diagrams that had higher tariffs given the opportunity cost for not attempting them at all, while it was less of a loss for not attempting items with diagrams with lower tariffs. However, further evidence would be needed to confirm this hypothesis.

Although it does not directly concern accessibility, our finding relating to the overall patterns of item omission across gender groups is noteworthy. Male candidates in the foundation tier papers had a higher tendency to omit items than their female peers did, and this was true for both items with and without diagrams. This finding corroborates the results from Matters and Burnett (1999)

indicating that the rate of omission was higher among male candidates than female candidates for constructed-response items.

It should be noted that this study was conducted using a limited dataset, based only on one exam series and a relatively small number of items involving diagram creation or augmentation. Despite this limitation, this research has demonstrated the potential of analysing omit rates to provide initial indications of item accessibility. However, in such analysis, other factors that can also influence omit rates, as discussed in this article, need to be kept in mind. A follow-up study involving a larger dataset would allow more fine-grained analyses of how the interactions between variables could potentially contribute to patterns of omit rates. Additionally, a larger dataset with more items with diagrams would enable further distinction between items involving diagram creation and those involving diagram augmentation. There could potentially be differences between items that require students to draw a diagram and items that require them to augment a partially provided one in terms of the nature and level of response strategy demand (see Pollitt et al., 2007). Items that require students to augment a partially provided diagram potentially pose less risk of a student being entirely unable to make an attempt as at least some of the diagram is already provided. It could be speculated that such differences may have implications for accessibility. In this study we could not further distinguish these two item types (create versus augment) as the numbers of items would have been even smaller. A larger dataset could allow examination of potential implications of these two item types for accessibility. Future studies should also consider gathering insights from students after they take an exam about why they leave out certain questions, to explore the contribution of different variables to omit rates and help establish the contribution of accessibility to patterns of omission.

Acknowledgement

The author would like to thank Jeff Heath for carrying out the coding of the items included in the study reported in this article.

References

Ainsworth, S., Prain, V., & Tytler, R. (2011). Drawing to learn in science. *Science, 333*(6046), 1096–1097.

Beauchamp, D., & Constantinou, F. (2020). Using corpus linguistics tools to identify instances of low linguistic accessibility in tests. *Research Matters: A Cambridge Assessment publication, 29,* 10–16.

Beddow, P. A. (2012). Accessibility theory for enhancing the validity of test results for students with special needs. *International Journal of Disability, Development and Education, 59*(1), 97–111.

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2013). Test accessibility: Item reviews and lessons learned from four state assessments. *Education Research International, 2013*(1), 952704.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*(1), 23–35.

Chang, H. Y., Lin, T. J., Lee, M. H., Lee, S. W. Y., Lin, T. C., Tan, A. L., & Tsai, C. C. (2020). A systematic review of trends and findings in research employing drawing assessment in science education. *Studies in Science Education, 56*(1), 77–110.

Clemens, N. H., Davis, J. L., Simmons, L. E., Oslund, E. L., & Simmons, D. C. (2015). Interpreting secondary students' performance on a timed, multiple-choice reading comprehension assessment: The prevalence and impact of non-attempted items. *Journal of Psychoeducational Assessment, 33*(2), 154–165.

Crisp, V., & Macinska, S. (2020). Accessibility in GCSE Science exams - Students' perspectives. *Research Matters: A Cambridge Assessment publication, 29,* 2–10.

Department for Education. (2015). *Biology, Chemistry and Physics GCSE subject content*.

LaDue, N. D., Libarkin, J. C., & Thomas, S. R. (2015). Visual representations on high school biology, chemistry, earth science, and physics assessments. *Journal of Science Education and Technology, 24,* 818–834.

Matters, G., & Burnett, P. C. (1999). Multiple-choice versus short-response items: Differences in omit behaviour. *Australian Journal of Education, 43*(2), 117–128.

OCR. (2018a). *GCSE (9–1) Gateway Science: Exploring our question papers*.

OCR. (2018b). *GCSE (9–1) Twenty First Century Science: Exploring our question papers*.

Ofqual. (2022). *Guidance on designing and developing accessible assessments: Consultation decisions*.

Papanastasiou, E. C. (2020). Can non-responses speak louder than words? Examining patterns of item non-response in TIMSS 2015. *International Journal of Quantitative Research in Education, 5*(2), 157–172.

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. Newton, J-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 166–206). Qualifications and Curriculum Authority.

Posit team. (2023). *RStudio: Integrated Development Environment for R* (Version 2023.12.0.369) [Computer software]. Posit Software, PBC. http://www.posit.co/

Prain, V., & Tytler, R. (2012). Learning through constructing representations in science: a framework of representational construction affordances. *International Journal of Science Education, 34*(17), 2751–2773.

Sarac, M., & Loken, E. (2023). Examining patterns of omitted responses in a large-scale English language proficiency test. *International Journal of Testing, 23*(1), 56–72.

Trumbo, J. (1999). Visual literacy and science communication. *Science Communication, 20*(4), 409–425.

Tytler, R., Prain, V., Aranda, G., Ferguson, J., & Gorur, R. (2020). Drawing to reason and learn in science. *Journal of Research in Science Teaching, 57*(2), 209–231.

Tytler, R., Prain, V., & Hubber, P. (2018). Representation construction as a core science disciplinary literacy. In K. S. Tang & K. Danielsson (Eds.), *Global developments in literacy research for science education* (pp. 301–317). Springer.

Unsworth, L., & Herrington, M. (2023). Visualization type and frequency in final year high school science examinations. *Research in Science Education, 53,* 707–725.

von Schrader, S., & Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education, 19*(1), 41–65.

Walland, E. (2024). Exploring speededness in pre-reform GCSEs (2009 to 2016). *Research Matters: A Cambridge University Press & Assessment publication, 37*, 57–73.

Wang, C., & Wei, B. (2024). Analysis of visual-based physics questions of the senior high school entrance examination in China. *Physical Review Physics Education Research, 20*(1), 010112.

How do candidates annotate items in paper-based maths and science exams?

Joanna Williamson (Research Division[1])

Introduction

Teachers, examiners and assessment experts know from experience that some candidates annotate exam questions. In this context, "annotation" includes anything the candidate writes or draws outside of the designated response space (the official answer space and "working out" space, and their margins). While many studies have analysed candidates' writing and drawing in response spaces, annotations are a potentially rich source of information about response behaviour that has been relatively overlooked. This article describes a study investigating candidate annotations in paper-based GCSE Combined Science and GCSE Mathematics exams. The motivation was to increase understanding of candidate response activity, in order to support the design of effective digital assessments in these subjects.

Candidates may annotate following explicit advice from teachers: commonly recommended exam strategies include the "BUG" technique (**b**ox the command word; **u**nderline key words; **g**lance to see if you've got all the info), for example, and the "HUA" method (**h**ighlight key words; **u**nderline command words; **a**nnotate).[2] In multiple-choice questions (MCQs), additional annotation in the form of marking or crossing out answer options can occur where students use elimination and guessing strategies. Annotation, including highlighting, is also of course recommended as a strategy to aid learning, and numerous studies have investigated the effect of annotation on comprehension in digital and paper-based reading (e.g., Ben-Yehudah & Eshet-Alkalai, 2018; Goodwin et al., 2020).

Some evidence on candidate annotation in exams has been captured by comparability studies investigating how responses and response behaviours change with test mode. These studies indicate that writing down "working out" and interacting with visuo-spatial information (e.g., graphs and diagrams) appears to matter for performance in maths and science assessments. Validity can be threatened when students cannot access "working out" space (Russell

1 Joanna conducted this research while working in the Research Division at Cambridge University Press & Assessment. She now works at Ofqual.

2 Examples of resources that recommend such techniques include materials by OCR (Butler, 2020), the Oxford Education Blog (Oxford Science Team, 2019), and BBC Bitesize revision guides.

et al., 2003), and although scratch paper can be provided, there are costs to transcription (Johnson & Green, 2006) and students may choose to work only in the mode in which the task is presented (Lemmo, 2023). Research findings suggest that students show their working out and annotate less frequently for digital items compared to their paper equivalents (Hughes et al., 2011; Johnson & Green, 2006). On highly visuo-spatial items (geometry, graphs) and items requiring annotation, students typically perform better on paper (Hughes et al., 2011; Keng et al., 2008; Lowrie & Logan, 2015). Lowrie and Logan (2015) also showed that students are more likely to use the provided diagram or graph to solve the item when working on paper.

Existing evidence on candidate annotations is not extensive and has several limitations. In the first place, it does not tell us much about the prevalence of annotation in candidate scripts — where rates of annotation are mentioned, it is typically to compare the on-screen and paper-based versions of one item. Furthermore, observations on annotation and reported rates of annotation tend to include all of candidates' "working out" — that is, including written response activity in designated "working out" spaces that was requested by the exam question. Due to the focus of the comparability studies, it is also the case that much of the evidence concerns items showing or expected to show mode effects. Finally, detailed studies on mode effects have often been smaller-scale qualitative studies and involved self-selecting samples of schools.

This article describes an exploratory study of candidate annotation that aimed to increase understanding of candidate behaviour when answering paper-based maths and science items. It used OCR's extensive script repository to gain insights from a wider range of schools and ability levels than considered in previous studies. The research questions investigated were:

1. Can candidates' script annotations be extracted at scale?
2. How often do GCSE candidates annotate their paper-based maths and science questions?
3. Does annotation rate vary by item type, subject or candidate grade?
4. What kinds of annotations do candidates make?

A particular motivation for this research was to support the design of effective digital assessments. Digital assessment has the potential to offer substantial benefits, but transitioning high-stakes maths and science assessments to on-screen formats also poses challenges (e.g., Ofqual, 2020). Response activities other than writing (e.g., problem solving, drawing, calculating) can require special characters and notation, special layouts, and the facility for candidates to freely express ideas and conduct "working out" (Williamson, 2023). These requirements can be difficult to fully accommodate in digital environments — at least in comparison with providing tools for drafting written English. Improving understanding of candidate response activity in paper-based exams could support effective digital assessment by helping assessment designers pinpoint aspects of maths and science response activity that may be impeded or supported by the affordances of a digital test environment. This can inform the design of digital-first maths and science items, and help identify how the response activity elicited by a paper-based item might change when the item is transferred to a digital format. It could also inform the design of digital platforms and on-screen tools.

Data

The research investigated annotations made by GCSE Combined Science and GCSE Mathematics candidates in June 2019. To do this, it analysed scans of handwritten exam scripts belonging to four random samples of 1000 candidates, one each from the Foundation (F) and Higher (H) tiers of each GCSE. The four random samples had very similar grade profiles to their respective full cohorts, as summarised in Table 1.[3]

Table 1: Summary statistics for the grades of the sampled candidates and their respective full cohorts

GCSE	Tier	Group	N	Mean grade[4]	Std Dev[4]	Median grade
Mathematics	F	Full cohort	28 005	3.1	1.3	3
		Sample	1000	3.2	1.2	3
Mathematics	H	Full cohort	16 948	6.3	1.6	6
		Sample	1000	6.1	1.6	6
Combined Science	F	Full cohort	10 175	5.2	2.4	3-3
		Sample	1000	5.2	2.4	3-3
Combined Science	H	Full cohort	6 794	10.0	3.4	6-5
		Sample	1000	9.9	3.4	6-5

The items selected for analysis are summarised in Table 2. Scanned script images were obtained for all items in Table 2, for the corresponding candidate samples in Table 1. In addition, random samples of 100 scanned script images (belonging to any candidates) were obtained for 24 of the items in Table 2, for use in training (see Methods section).

The GCSE Mathematics exams did not feature MCQs, but it was possible to include a range of MCQs from GCSE Combined Science. The selected items were the first and last two MCQs from alternate GCSE Combined Science papers. The reason for this choice was to:

- include both easier and harder multiple-choice items, for both tiers
- analyse Foundation and Higher tier responses to the same items (the final two MCQs of the Foundation paper are typically also included as the first two of the corresponding Higher tier paper)
- avoid further selection effects by manually choosing specific items.

3 The data used in this research was collected as part of the usual marking and processing of candidates' examination scripts. Data has been stored and used in line with Cambridge University Press & Assessment's Data Privacy notice (https://www.cambridge.org/legal/candidate-privacy-notice).

4 GCSE Combined Science is a double award GCSE in which candidates study all three sciences (Biology, Chemistry and Physics). To reflect the larger qualification size, candidates receive a double GCSE grade consisting of two identical or adjacent numerical grades, from 9-9 (the highest grade) to 1-1 (the lowest grade). For the purposes of calculation, all candidates were assigned a numerical grade equivalent. Grades X, U and candidates with no result were assigned the grade value zero. GCSE Mathematics grades 9 to 1 were given their face value (i.e., 9=9, 8=8, ... 1=1). GCSE Combined Science grades were assigned values 1 to 17 as follows: 9-9 = 17, 9-8 = 16, ... 1-1 = 1.

Non-multiple-choice items were chosen to include both low- and high-tariff items, a range of topic areas, and items with different features (e.g., graphs, diagrams, tables, equations, calculations). Graphics tasks are defined as items containing "high concentrations of visual-spatial information, including graphs, maps and diagrams" (Lowrie & Logan, 2015, p. 650).

Table 2: Summary of items analysed

GCSE	Total	MCQ		Graphics task		Calculation required	
		Yes	No	Yes	No	Yes	No
Combined Science – Foundation							
Biology	5	4	1	3	2	1	4
Chemistry	6	4	2	1	5	2	4
Physics	6	4	2	4	2	3	3
Subtotal	17	12	5	8	9	6	11
Combined Science – Higher							
Biology	6	4	2	3	3	2	4
Chemistry	6	4	2	1	5	2	4
Physics	6	4	2	4	2	3	3
Subtotal	18	12	6	8	10	7	11
Mathematics – Foundation	6	0	6	3	3	5	1
Mathematics – Higher	6	0	6	3	3	6	0
Total	**47**	**24**	**23**	**22**	**25**	**24**	**23**

The items selected were not a representative sample of all GCSE Mathematics and Combined Science items, because some items were excluded a priori due to the response space or working out space being integrated into the question. This occurs for example where candidates are invited to "Complete this table..." or "Show on the grid below...". Deciding which candidate markings should be classified as annotations would have been very arbitrary for these items, hence they were excluded.

Two items originally selected for analysis were later replaced. Both included a graph with a fine-grained grid, and the method developed for isolating candidate annotations was not able to reliably extract candidates' annotations from the grid.

Methods

Extracting annotations from script images

The first step was to develop a method for extracting candidate annotations from a scanned exam script. This was achieved using image processing techniques; all image processing and machine learning in subsequent steps was carried out in Python for speed and to make use of the libraries OpenCV and scikit-image.[5]

The annotation extraction algorithm takes the following inputs: a file path for the candidate's full scanned script, a file path for an unmarked copy of the exam

5 OpenCV and scikit-image are large and well-known Python libraries containing functions for image processing and computer vision tasks.

paper to serve as the reference image, and the page reference and coordinates for the area(s) of interest on the reference image. The algorithm applies these steps:

1. Selects the page and area of interest from the scanned script and aligns it to the reference image.
2. Applies a sequence of image adjustments including blurring, thresholding, dilation, and erosion to the aligned target image. The goal is to make any candidate annotations prominent, while reducing flecks, spots, and creases in the target image that could be mistaken for annotations. The identical sequence of image adjustments is applied to the reference image.
3. Subtracts the adjusted reference image from the adjusted target image.
4. Applies a further sequence of image adjustments to the remaining image, which consists solely of (adjusted) annotations.
5. Reduces the annotations to a set of features (quantitative variables) including number of remaining objects (pixel clusters), and number of remaining objects exceeding the size of a typical hand-written letter or number.

Training a classification algorithm

The next step was to train a machine learning algorithm to classify new item images as annotated or not annotated.[6] To produce training data, the random samples of 100 script images were processed using the annotation extraction algorithm, for 24 items. The 24 items selected were a subset of those in Table 2, chosen to include a range of item types (e.g., MCQ and non-MCQ, items both with and without graphs and diagrams). This processing resulted in a dataset of features (quantitative variables) for 2400 item-level script images. A variable for presence of annotation was manually added to label each of these 2400 images as annotated or not annotated. This was not too time-consuming, since the dataset could be sorted by extracted features (e.g., numbers of pixel clusters) and the images could generally then be quickly identified as annotated or not (with rapid viewing of the images to confirm, rather than determine, the correct labelling). Some script images required careful scrutiny to inform whether they should be classified as annotated or not annotated.

The labelled dataset (for 2400 item images) was then split into training (70 per cent) and testing datasets (30 per cent) and used to train several simple classification algorithms. The final choice of classification algorithm was an XGBoost algorithm[7] trained on the following annotation features:

1. S01-S07: the size (in pixels) of the seven largest distinct objects
2. Count: the number of distinct objects with size at least 500 pixels
3. Count_SSI: the number of distinct objects with size at least 500 pixels, in a region defined as special interest (e.g., the graph or diagram, if one exists)
4. Count_safe: the number of distinct objects with size at least 1500 pixels.

6 A simpler approach tried first was to compare the total number of black pixels in scanned item images with the total in the reference image. This was not successful, because variation in the scanned item images (e.g., page creases, unexplained speckling) masked the "signal" of annotations.

7 An XGBoost algorithm uses gradient-boosted decision trees to solve supervised machine learning problems (in this case, a classification task).

The challenge in developing the annotation extraction and classification methods was to successfully identify annotations while avoiding false positives. The features of script images that caused the most difficulties were scans with many page imperfections (e.g., creases, speckling), particularly in combination with minimal annotations, and the fact that some annotations appeared only faintly when scanned — perhaps due to the candidate's pen or pencil.

Main processing

The annotation extraction algorithm was applied to scanned script images for all items in Table 2, for the corresponding candidate samples in Table 1. Each item image was then classified as annotated or not annotated using the classification algorithm.

After classifying all item images, analyses were carried out based on simple descriptive statistics. For each item, the item annotation rate was calculated as the percentage of the 1000 candidates sampled who annotated that item.

A separate set of annotation rates was also calculated, considering only item images from candidates who attempted the item (i.e., where the examiner recorded a mark, even if zero). The omit rates for the items in this study were generally very low, however, so most items showed no difference in the calculated annotation rate. For the few items where there was a difference, the "attempts only" annotation rate was always higher by 1 to 5 percentage points. For simplicity, the results in this article report only the overall annotation rates (i.e., considering all item images, whether the candidate attempted the item or not), the lower of the two estimates. The "attempts only" annotation rates are included in Table A1 in the Appendix.

Annotation heat maps

To look at patterns of annotation, the candidate annotations for each item were combined and overlaid onto the reference image to create "heat maps". These graphs use colour intensity to indicate how frequently each pixel was annotated. Areas of the item annotated frequently show more intense colour, while areas not annotated at all show only the black and white reference image.

Isolating the annotations to create the heat maps required similar steps to those in the annotation extraction algorithm, but the exact sequence of image adjustments applied was different due to their different purposes. Rather than helping the annotations appear prominent (for the purposes of classification), the goal in the context of creating the heat maps was to preserve as much detail of candidates' annotations as possible, while still removing the reference image. A consequence of the lighter-touch image adjustments was that very faint images of questions printed on the reverse of the page were sometimes preserved along with candidates' annotations. These images of questions printed on the reverse were not visible as marks when looking at a single image (if traces were visible at all, they appeared as a more shadowy area of the white space) and consequently they were not captured by the annotation extraction algorithm. Hence, they did not affect the classification of images as annotated or not annotated (which was determined using the results of the annotation extraction algorithm) or the subsequent calculation of annotation rates. However, when

the annotation images from many candidates were combined to create the heat maps, the images of some questions became visible — since the same very faint image had been captured by the heat map algorithm in precisely the same location in all candidate annotation images. This was particularly noticeable for item M08 (Figure 2). The annotation rate for this item was 93 per cent, so the heat map represents the combined images of around 930 candidates, and the graph question from the reverse of the page can be seen on the right of the image. It is important to emphasise that this visual effect in the heat maps did not impact the calculation of annotation rates.

Results

Extraction of annotations and classification

The processes described in the methods section were able to extract annotations from script images at scale, and the final classification algorithm was able to classify new item images as annotated or not annotated. In particular, the algorithm was able to classify items not included in the training data.

The classification algorithm demonstrated good accuracy (Table 3). The variable "largest distinct object size" (S01) was by far the most important feature for the final classification algorithm.

Table 3: Metrics for the final classification algorithm, based on accuracy of classification of images in the testing dataset compared to manual coding of these images

Metric	Definition	Value
Accuracy	Proportion of total classifications that were correct	0.967
Precision	Proportion of total positive classifications that were true positives	0.942
Recall	True positive classifications as a proportion of total actual positive instances	0.968
F1 score	Harmonic mean of precision and recall	0.955

How often did candidates annotate items?

The overall rate of annotation across all items and candidates in this study was 40 per cent. The least annotated item was a multiple-choice Chemistry question (Figure 1) which was annotated by 8 per cent of sampled Foundation tier GCSE Combined Science candidates. The most frequently annotated item was a Higher tier GCSE Mathematics question (Figure 2) asking students to calculate the perimeter of a compound shape, which was annotated by 93 per cent of candidates. The annotation rates for all items in the study are listed in the Appendix (Table A1).

1 Which of these processes is an example of a **physical** change?

A Combustion

B Freezing

C Neutralisation

D Oxidation

Figure 1: Heat map showing annotation on item C01 (Chemistry, Foundation tier)

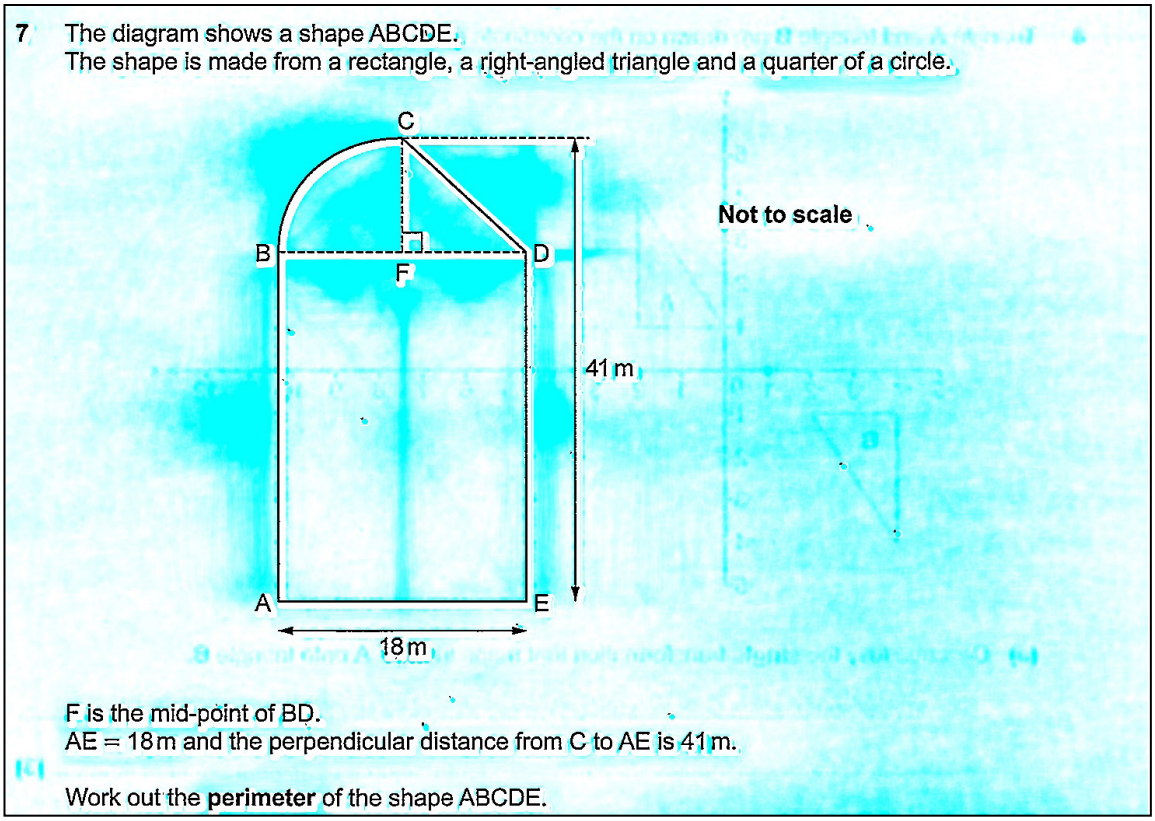


Figure 2: Heat map showing annotation on item M08 (Maths, Higher tier)

Annotation rates by item features

For context, Figure 3 shows the distributions of item-level annotation rates by GCSE, tier and subject area. For Higher tier GCSE Combined Science, the mean annotation rate was slightly higher for Chemistry items than for Physics and Biology items, whereas for Foundation tier, slightly higher rates of annotation were found for Physics and Chemistry items than Biology items. The Higher tier GCSE Mathematics items tended to be annotated most frequently, but this may simply reflect the particular items sampled and should not be over-interpreted.

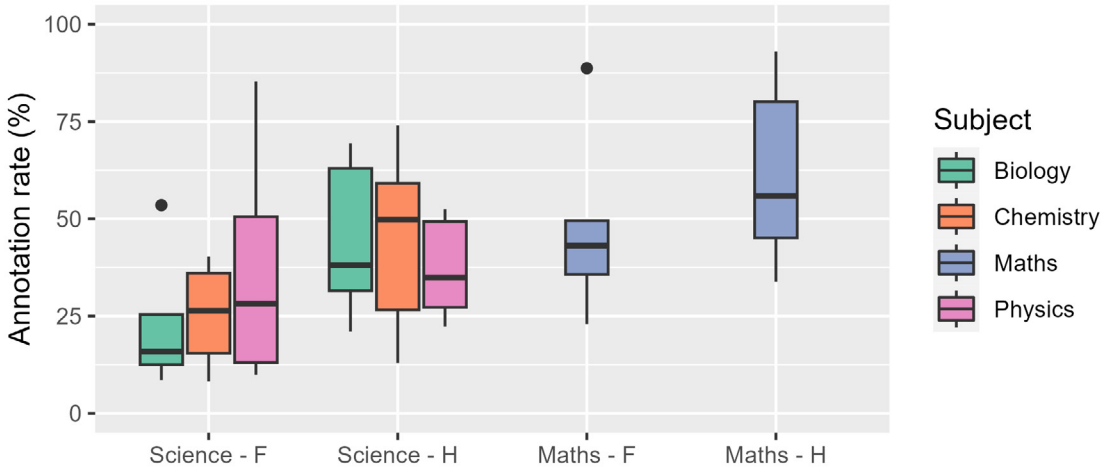


Figure 3: Annotation rates by GCSE, tier and subject area

The more interesting comparisons are those comparing rates of annotation for items taken by the same sample of candidates. Figure 4a shows that for all four candidate samples, candidates annotated graphics tasks more frequently than other items. When items featuring tables and equations were also grouped together with graphics tasks, the pattern became even more pronounced (Figure 4b).

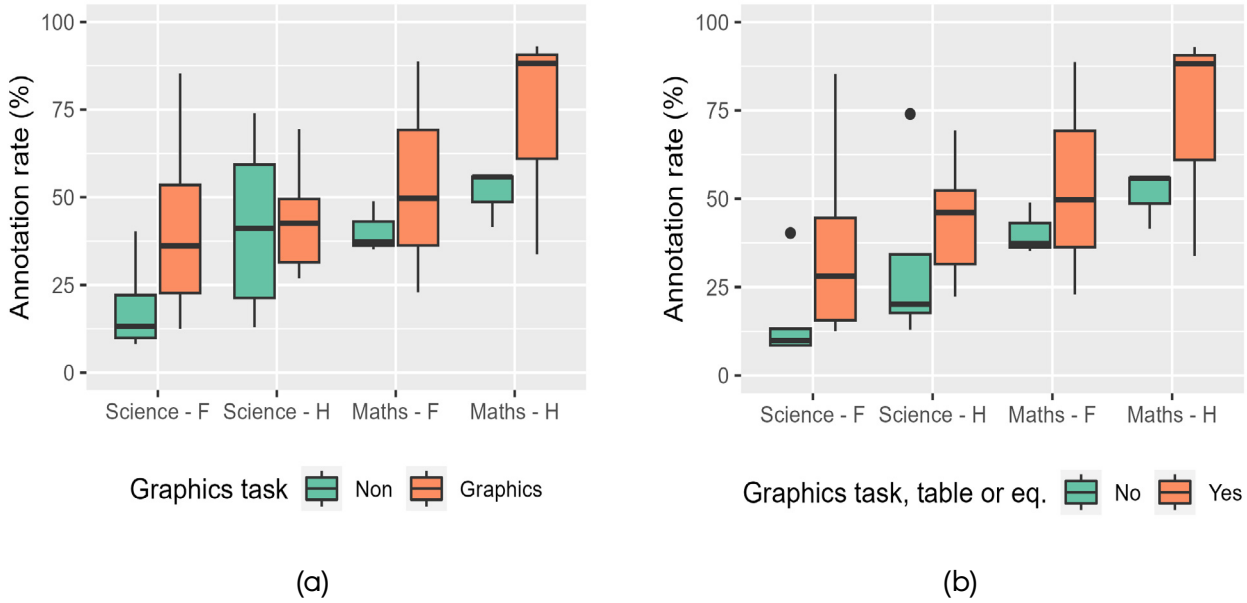


Figure 4: Annotation rates for graphics (including or excluding tables and equations) and non-graphics tasks

Another item feature that was expected to be associated with candidate annotations was a requirement for calculation. Figure 5 shows that items requiring calculation were indeed annotated more frequently than items not requiring calculation. Furthermore, within both categories, graphics tasks (including items with tables and equations) were still generally annotated more frequently than non-graphics tasks. This was not the case for the Higher tier Combined Science items, and this may reflect other item characteristics not accounted for.

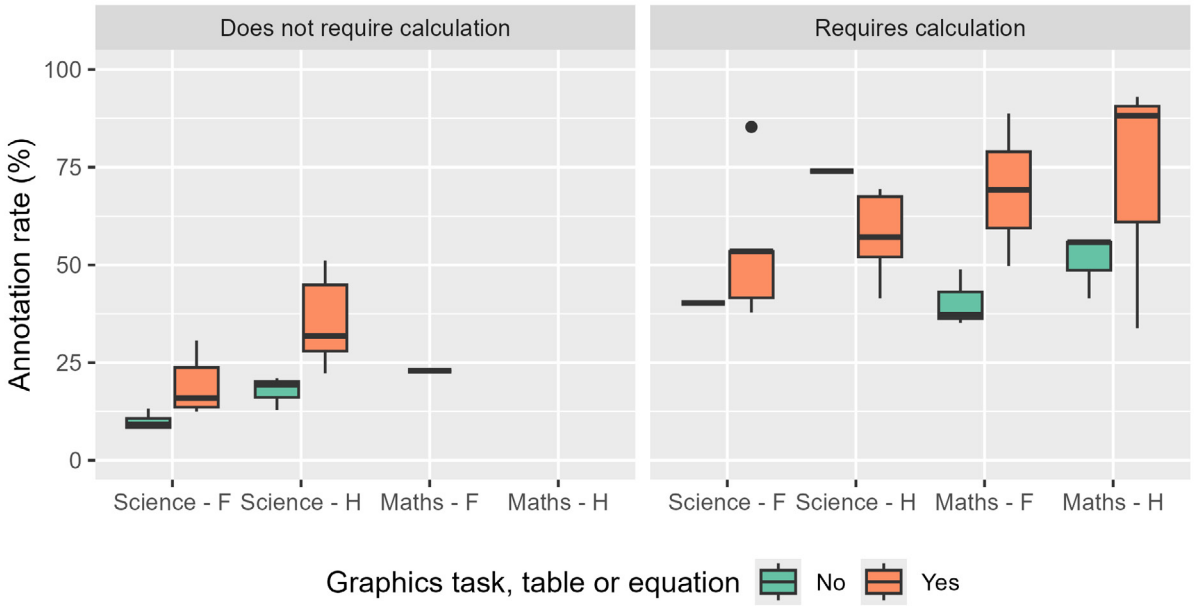


Figure 5: Annotation rates by graphics task (including items with a table or equation) and requirement for calculation

Finally, Figure 6 shows the distributions of annotation rates for MCQs compared to other items.[8] Three points are worth noting from Figure 6. Firstly, the annotation rates for MCQs spanned a wide range. Secondly, the annotation rates for MCQs that were graphics tasks or required calculation were noticeably higher than the annotation rates for other MCQs, in line with the pattern seen for items overall. And thirdly, the annotation rates for Higher tier MCQs that were graphics tasks or required calculation were comparable to the annotation rates for non-MCQs with these features, in both GCSE Combined Science and GCSE Mathematics.

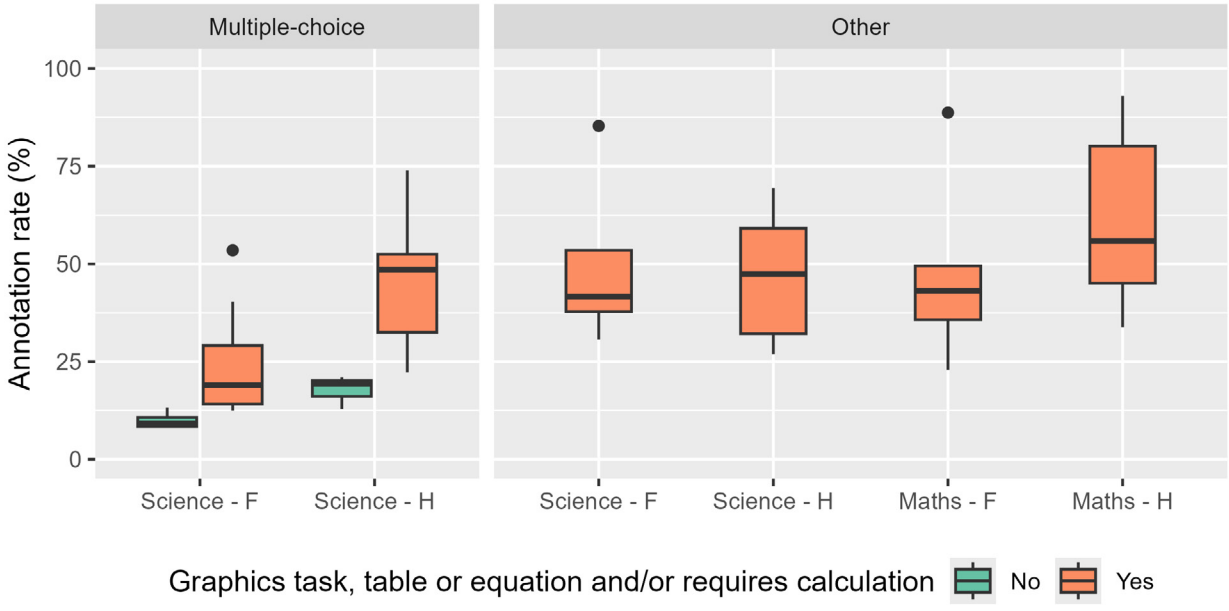


Figure 6: Annotation rates by question type (MCQ or non-MCQ), graphics task (including items with table or equation) and requirement for calculation

8 Note that among the "Other" (non-MCQ) items, all items were either a graphics task, featured a table or equation, or required calculation.

Annotation rates by grade and tier

Across all four subject areas, candidates with higher grades in the relevant GCSE tended to annotate items more frequently. Figure 7 shows the percentage of item images that were annotated in each subject area, by grade in the relevant GCSE (i.e., GCSE Mathematics grade for the maths items, and GCSE Combined Science grade[9] for the biology, chemistry and physics items).

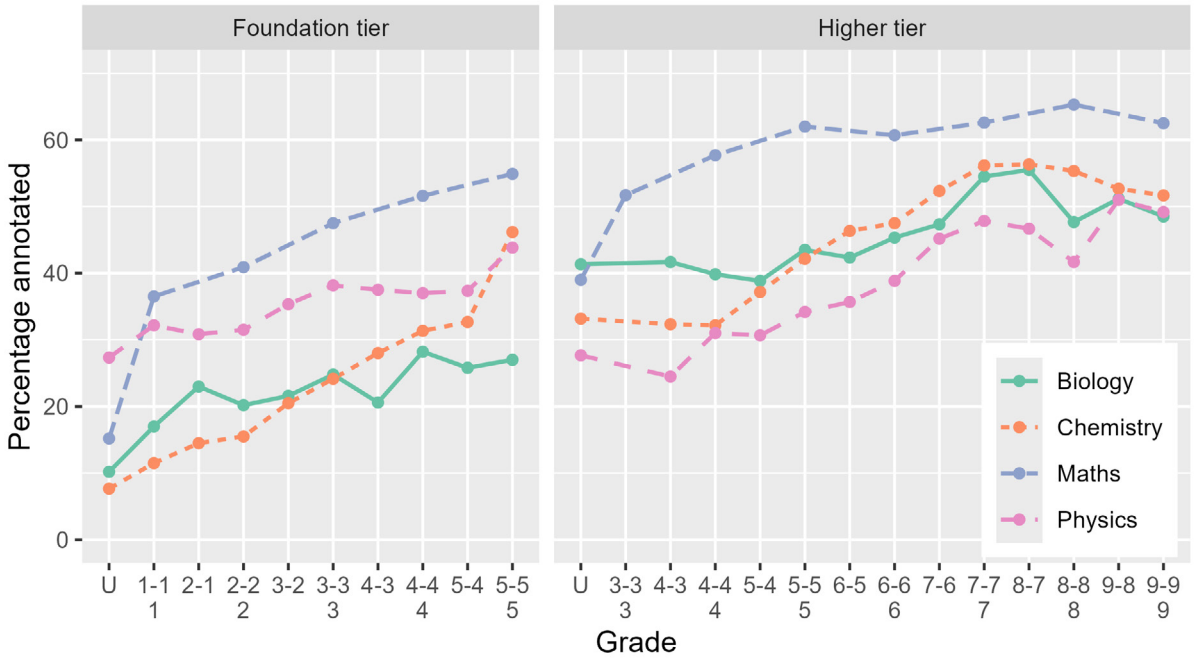


Figure 7: Percentage of item images annotated by relevant GCSE grade

For items that appeared on both Foundation tier and Higher tier papers, the rate of annotation was higher among Higher tier candidates for all items except P04 (Table 4). The largest difference was 20 percentage points, for the GCSE Mathematics item M03.

9 As described earlier, GCSE Combined Science candidates are awarded a double grade on the scale 9-9 to 1-1, which represents their achievement across all three of the science subjects.

Table 4: Annotation rates for items on both Foundation and Higher tier papers

Item	Annotation rate (%)		Difference
	Foundation tier	Higher tier	
B03	25.4	32.5	7.1
B04	15.9	31.2	15.3
B09	53.5	69.4	15.9
C04	13.2	19.3	6.1
C09	37.8	51.1	13.3
M02	37.3	41.5	4.2
M03	35.2	55.8	20.6
P03	12.5	22.3	9.8
P04	53.5	41.5	-12.0

Types of annotation observed

Several different types of annotation were observed in candidates' script images. This section briefly describes each category and illustrates with examples.

1 Highlighting key information

Candidate annotations included underlining, circling and boxing of key words and values in the question text. This was seen across multiple items, including the item in Figure 1 where the heat map indicates annotation of the key word "physical". Annotation of key words can also be seen in the heat maps in Figure 8, Figure 9, Figure 13 and Figure 22, and examples of individual candidates' annotations of this type can be seen in Figure 12, Figure 23 and Figure 25.

2 Crossing/ticking multiple-choice options

For many MCQs, the heat map revealed annotation of the answer option labels and at the ends of answer options, as shown in Figure 8 and Figure 9. Although the heat maps indicate where annotation occurred, it is often necessary to look at individual scripts to determine exactly what marks individual candidates made. Figure 10 shows an example of one candidate's actual annotations — in this case, small crosses at the end of three answer options. Neither of these MCQs require calculation, and they offer answer options in the form of parallel statements.

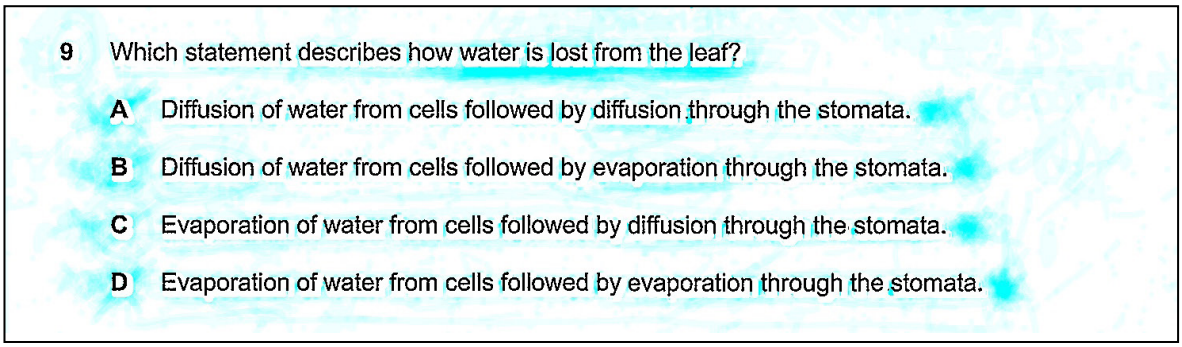


Figure 8: Heat map showing annotation on item B05 (Biology, Higher tier)

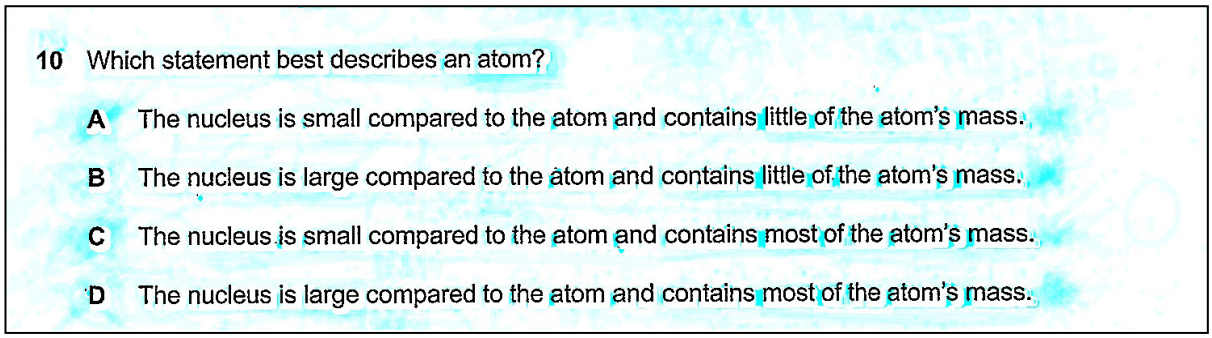


Figure 9: Heat map showing annotation on item C04-F (Chemistry, Foundation tier)

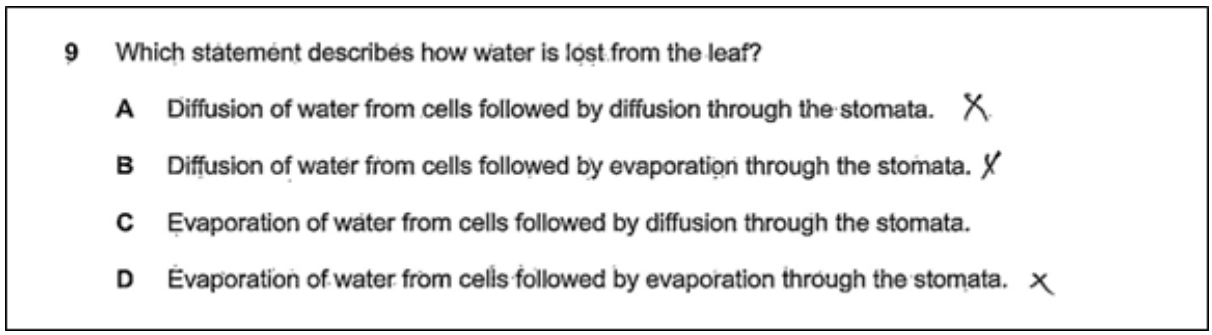


Figure 10: Example of candidate annotations marking crosses on three answer options, B05 (Biology, Higher tier)

3 Annotating the question with related facts or rules

Candidate annotations included candidates writing down rules, notes and facts related to the question. Figure 11, for example, shows candidate annotation including the SOHCAHTOA[10] mnemonic, and Figure 12 shows where a candidate has underlined "exothermic" and added the annotation "gives out heat", along with "oxidation is loss".

10 SOHCAHTOA is a mnemonic for the definition of trigonometric functions. For angle θ in a right-angled triangle, the trigonometric functions are defined in terms of the ratios of sides: **s**ine θ = **o**pposite/**h**ypotenuse, **c**osine θ = **a**djacent/**h**ypotenuse, and **t**angent θ = **o**pposite/**a**djacent.

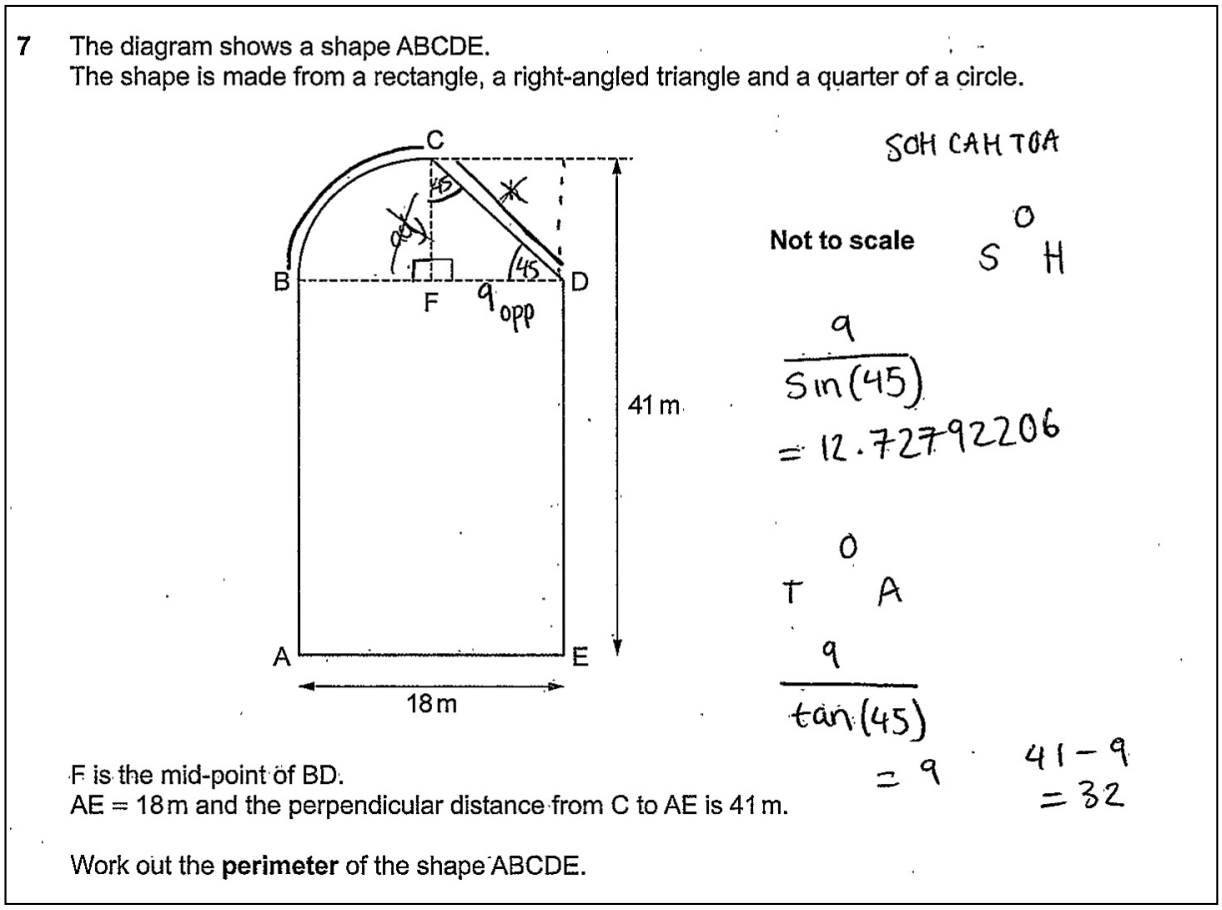


Figure 11: Example of candidate annotations noting angle facts, M08 (Maths, Higher tier)

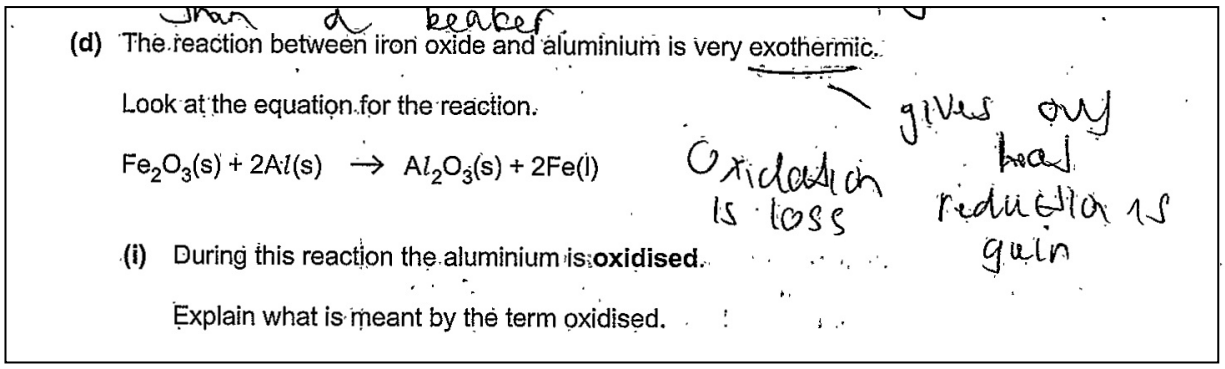


Figure 12: Example of candidate annotations noting information relating to the terms "exothermic" and "oxidised", C09 (Chemistry, Higher tier)

4 Annotating a graph or figure

The graphs and figures included in items were frequently annotated by candidates, for example by using values provided in the question text. The heat map in Figure 13 shows blue horizontal and vertical lines indicating where multiple candidates marked key positions or values on the graph as part of their working out.

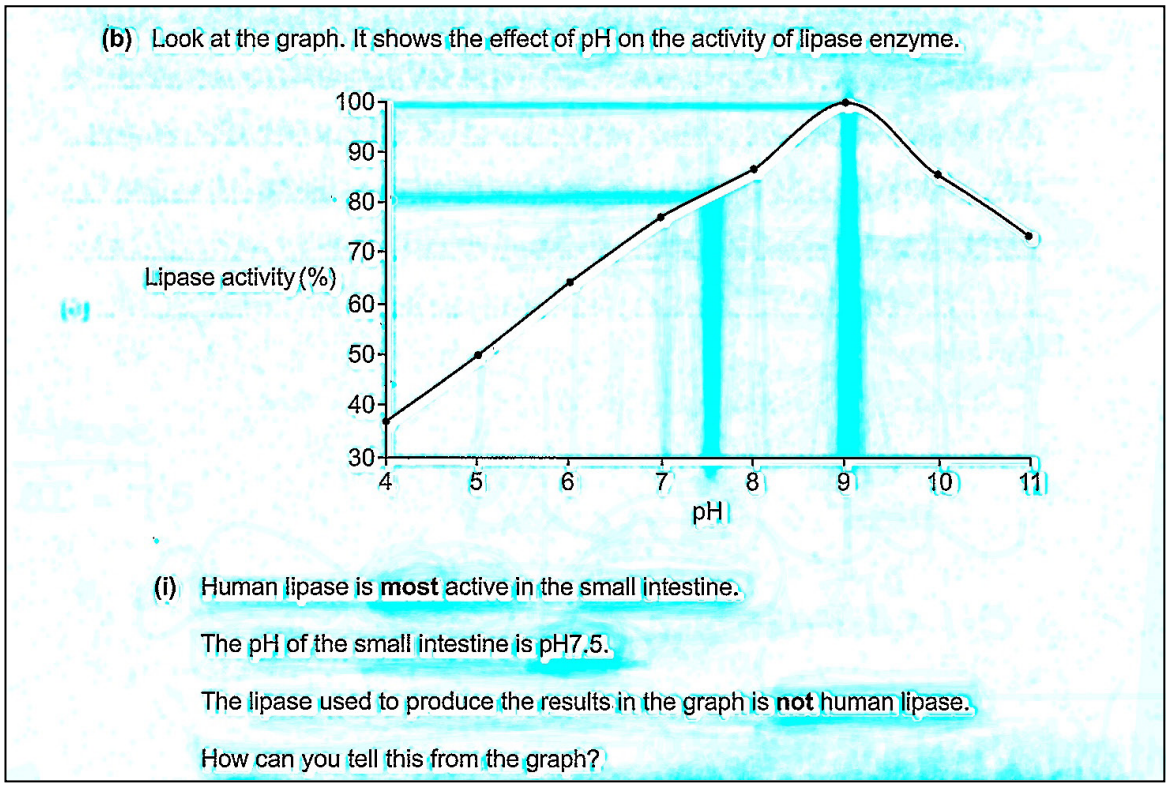


Figure 13: Heat map showing annotation on item B08 (Biology, Higher tier)

Figure 14 shows that many candidates annotated areas corresponding to angles on the diagram. In particular, the blue areas suggest frequent annotation of the angles that can be deduced using the "alternate angles" rule within parallel lines, the knowledge that angles on a line add up to 180°, and the knowledge that angles inside a triangle add up to 180°. As an example, Figure 15 shows where one candidate has added the values of four angles that can be deduced using these rules. Figure 16 shows where another candidate has drawn a "Z" onto the diagram, perhaps to confirm or identify where the alternate angles rule may help.

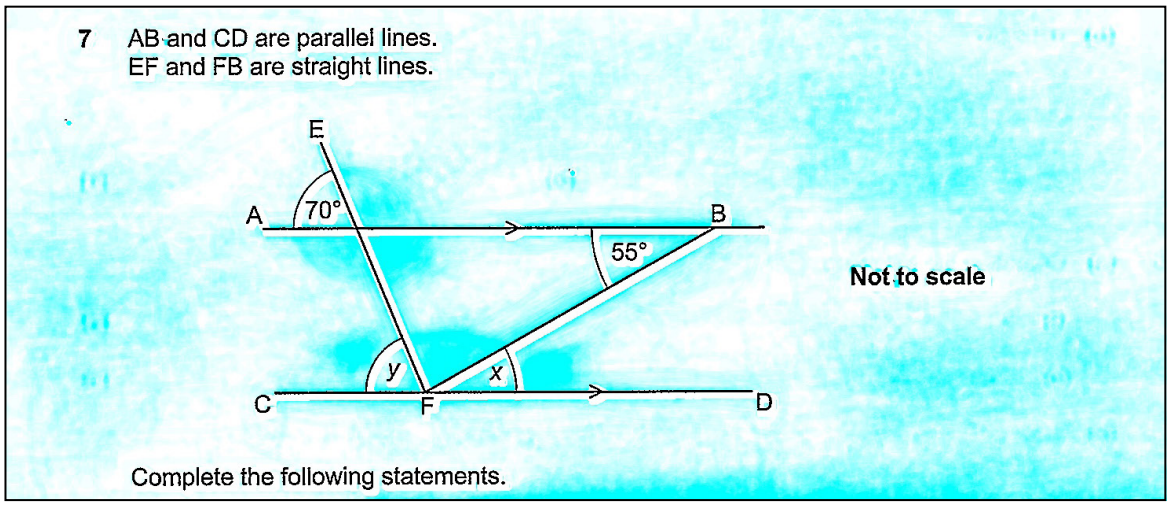


Figure 14: Heat map showing annotation on item M01 (Maths, Foundation tier)

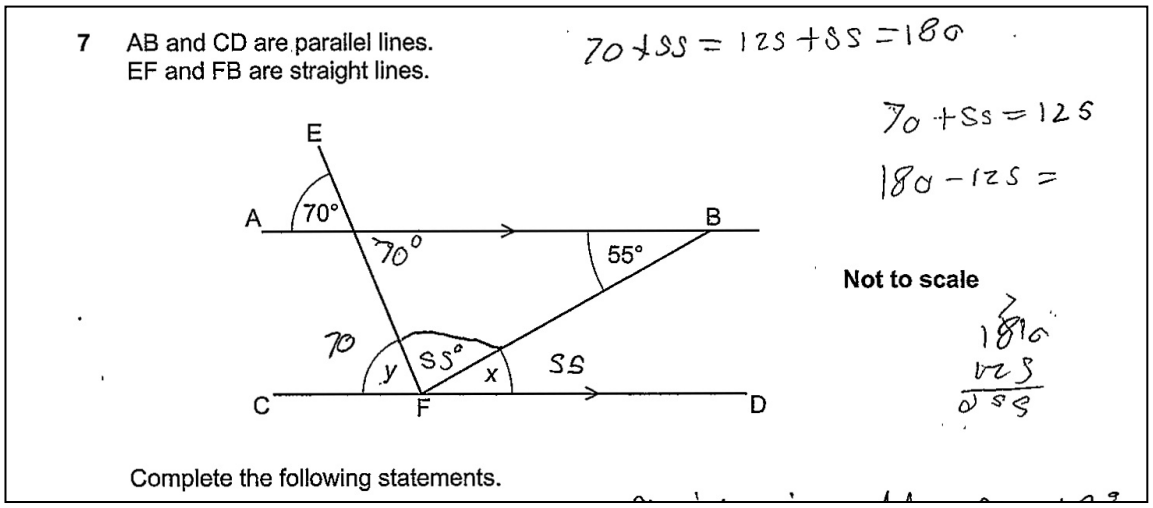


Figure 15: Example of candidate annotations suggesting use of several rules regarding angles, M01 (Maths, Foundation tier)

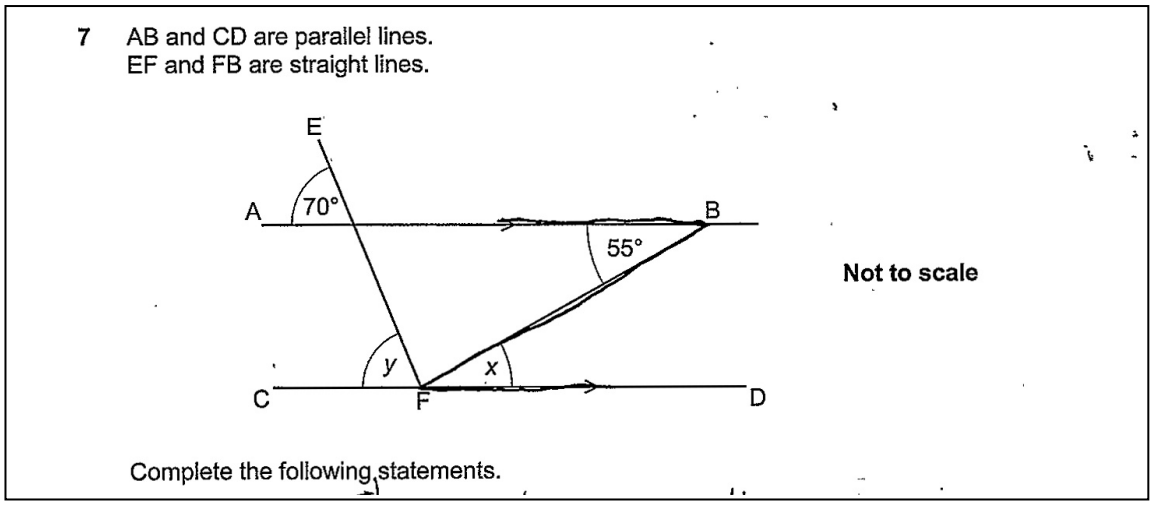


Figure 16: Example of candidate annotations suggesting use of "alternate angles" rule, M01 (Maths, Foundation tier)

5 "Working out" in or immediately around the question text

Inspection of script images and the heat maps shows clearly that candidates carried out "working out" on or directly around question text even when space was available elsewhere (e.g., in a designated "working out" space, or in white space on the page). A hypothesised explanation for this is that candidates might value the immediacy of writing onto the question text, and perceive a lower risk of slips or loss of attention, compared to working out in the designated answer space. By writing onto the question itself, candidates can use the information from the question while avoiding the effort and risk of copying values to a new area of the page. Figure 11 (shown earlier) illustrated this kind of annotation, with perimeter calculations written immediately next to the diagram and question text rather than in the working out space.

Figure 17 shows an item where working out in or immediately around the question text was particularly prevalent. The dense blue areas around the sequence numbers show that extensive annotation occurred here, and Figure 18 and

Figure 19 show examples of candidate annotations that contributed to this heat map pattern. The same pattern of annotation also occurred for the other mathematical sequence item in the study (Figure 20 and Figure 21). For both sequence items, it appears candidates used the spatial layout of the question text to structure their working.

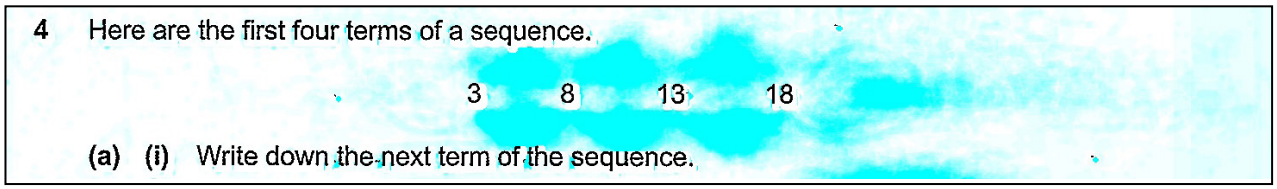


Figure 17: Heat map showing annotation on item M06 (Maths, Foundation tier)

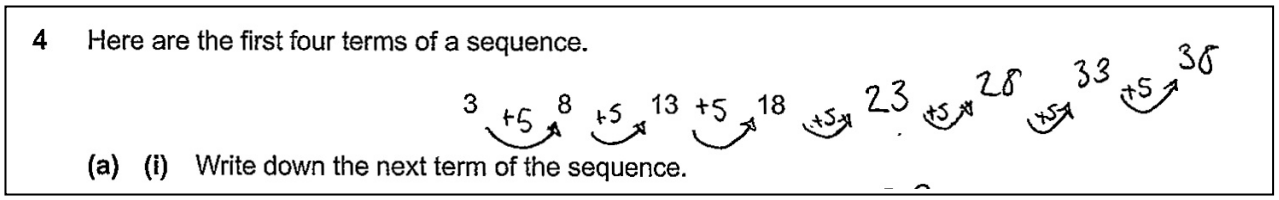


Figure 18: Example of candidate annotations showing addition between terms and sequence continuation, M06 (Maths, Foundation tier)

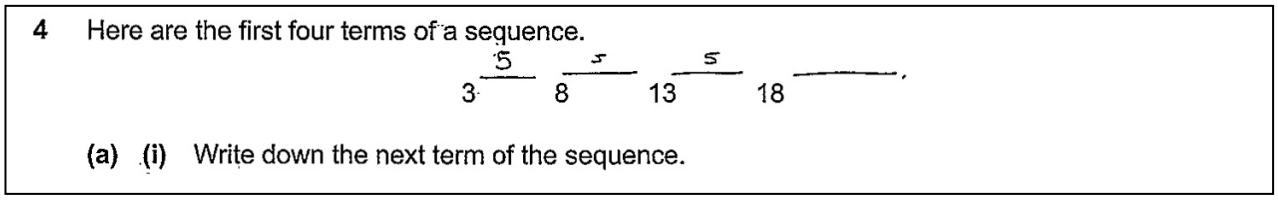


Figure 19: Example of candidate annotations marking on the differences between sequence terms, M06 (Maths, Foundation tier)

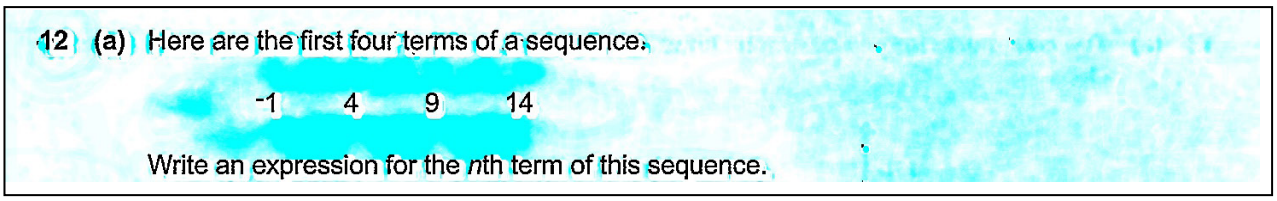


Figure 20: Heat map showing annotation on item M09 (Maths, Higher tier)

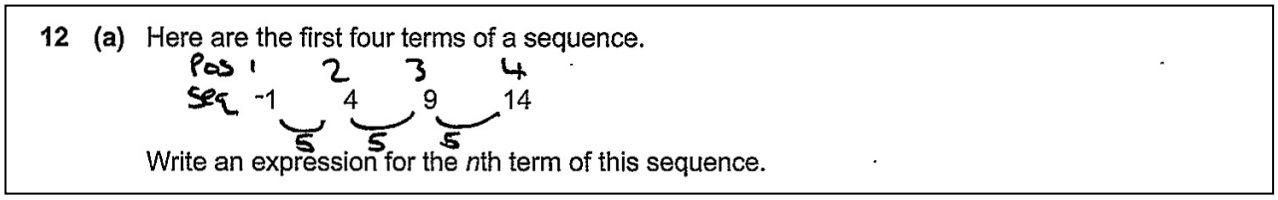


Figure 21: Example of candidate annotations numbering the sequence terms and marking on the differences between terms, M09 (Maths, Higher tier)

6 "Working out" where no space is explicitly provided

For several MCQs, candidates were asked to calculate values, but no "working out" space was provided — the only designated response space was a box for the letter of the answer option. For these MCQs, candidates unsurprisingly made use of the white space next to the answer options to carry out calculations and sometimes sketching. Figure 22 shows the heat map for an item where this was common, and Figure 23 shows an example of one candidate's annotations.

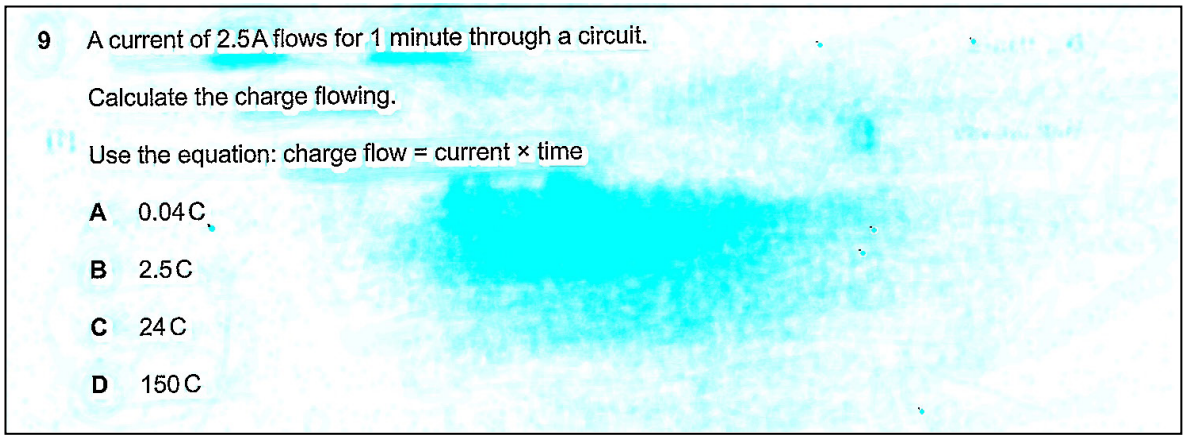


Figure 22: Heat map showing annotation on item P05 (Physics, Higher tier)

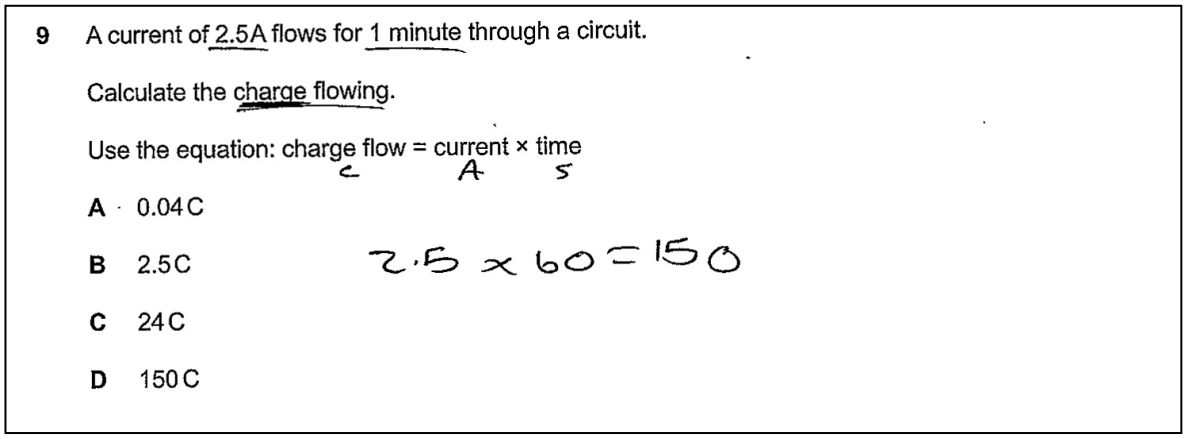


Figure 23: Example of candidate annotations including underlining of key information and calculation in white space, P05 (Physics, Higher tier)

7 "Overspill" working out

Finally, on some items, the heat map suggests that many candidate annotations were part of extensive working out that did not fit into the designated response space. Figure 24 shows the heat map for an item where this was common, and Figure 25 shows an example of one candidate's actual annotations (the figure shows only the area around the question text — the remainder of this candidate's working out was in the designed response space and is not shown).

Figure 24: Heatmap showing annotation on item M03-H (Maths, Higher tier)

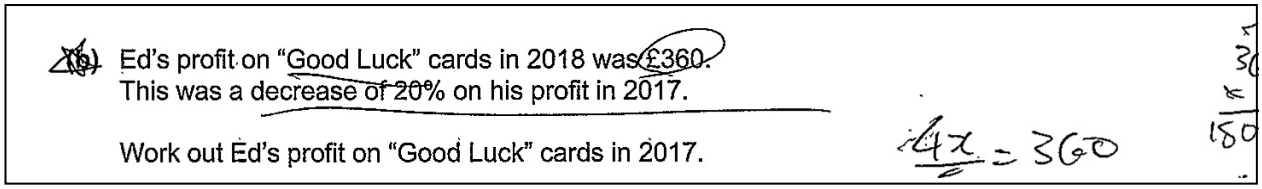


Figure 25: Example of candidate annotations including "overspill" from designated response space, M03-H (Maths, Higher tier)

Discussion

This exploratory project showed that candidate annotations can be extracted at scale from exam scripts, and that annotation rates can be calculated quickly for large samples of candidates. The heat map representations were able to reveal which areas of an item candidates tended to annotate more or less frequently, sometimes highlighting strong patterns in candidate response behaviour. This was supported by inspection of example script images to better understand the nature of individuals' annotations.

The GCSE Mathematics and Combined Science items sampled for this project were annotated fairly frequently. The overall rate (across all items and candidates) was 40 per cent, and the annotation rate for individual items ranged from 8 to 93 per cent. In general, higher-attaining candidates annotated the items at higher rates than lower-attaining candidates. For items that appeared on both Foundation tier and Higher tier papers, the Higher tier candidates almost always annotated that item at a higher rate. The current study did not attempt to evaluate the usefulness of candidates' annotations in terms of helping them successfully answer specific items, but it is interesting to reflect on the variation in annotation rates across grades in light of the work by Hughes et al. (2011). Their study found larger mode effects for higher-attaining students on graphics tasks and items requiring working, which the authors interpreted as higher-attaining students being less able (in the on-screen test mode) to use their preferred strategies of jotting and annotating.

In terms of variations across item types, the results were in line with expectations based on the research literature. Items with high concentrations of visuo-spatial information including graphs and diagrams were annotated more frequently than items without these features. In addition, items that required candidates to carry out calculation were annotated more frequently than items that did not require calculation, even when a dedicated "working out" space for this calculation was provided to candidates. As noted in the results, a hypothesised explanation for this is that candidates may perceive a benefit from the immediacy of working directly alongside information presented in the question text. This idea extends Johnson and Green's (2006) reflections on the role of proximity when candidates respond to maths items; they theorised that a greater distance between information presented and the working out space (e.g., between an on-screen question and scratch paper) could be a cause of transcription difficulties, which introduce errors into candidates' responses.

While graphics tasks and items requiring calculation were more frequently annotated, a substantial minority of candidates also annotated other item types. The annotations found on these items included highlighting key information in the question, jotting down facts or rules, and marking or eliminating multiple-choice answer options. The annotations candidates made to key words and values were often clearly visible in the heat maps, indicating that large number of candidates had chosen to annotate the same parts of the question text. This annotation behaviour appears in line with known exam techniques (such as "BUG" and "HUA") mentioned at the start of this article that encourage candidates to mark the key information in questions in order to aid accurate reading and responding. The annotations candidates made to multiple-choice answer options were again a form of annotation anticipated from the literature on MCQ response behaviour. The presence of similar marks on three out of four answer labels or answer rows suggests elimination, but this could be consistent with various response behaviours: for example, a step-by-step process that eliminates answers one by one, or a confirmatory elimination that checks off incorrect answer options after using another strategy to determine the correct answer.

The MCQs analysed in this work showed a range of candidate annotations and annotation rates. Most notably, the MCQs that were graphics tasks or required calculation were annotated at comparable rates to the non-MCQ items with these features. MCQs are typically considered less challenging to implement in digital modes than constructed response items (e.g., Crisp & Ireland, 2022; Drijvers, 2019). The response space (e.g., checkbox) can remain "the same" in a digital format, and MCQs avoid the need to input special characters or formats, and input or transcribe working out. However, the results relating to MCQs underline the broader point made by this research, which is that focusing solely or mainly on designated response spaces may risk overlooking what candidates are doing or producing on their way to that response.

A key limitation of this research is that the method developed is not suitable for all maths and science questions, because it relies on defining areas of the page as "response space" (and correspondingly, "not a response space"). Questions where a response space is fully integrated into the question text or stimulus cannot be analysed with this method, and for this reason, the research could not analyse a representative sample of all maths and science items.

While acknowledging this limitation, the research has provided evidence for patterns of annotation across a wide range of maths and science items, and the types and rates of annotation found suggest that this aspect of candidate response behaviour merits attention. Understanding response activity is important for assessment validity, and it is hoped that the evidence from this research can help inform the development of high-quality digital assessment in maths and science. It could help identify response behaviours that may be impeded or supported by the affordances of a digital test environment, and help anticipate how the response activity elicited by a paper-based item might change when the item is presented in a digital format.

This exploratory study could be followed up by further work in several areas. These include: developing reliable categorisations of the different annotations

observed; comparing patterns of annotation to the cognitive activity that items are designed to require or elicit; investigating the importance or value of specific annotations; investigating whether individual candidates demonstrate consistent annotation behaviours across items; and, relatedly, investigating whether there is a relationship between annotation behaviour and teaching and learning. The method of extracting and summarising candidate annotations could be applied to written exam papers in other subjects (e.g., English Literature). The approach requires access to large volumes of script images, but is otherwise quick and low-cost, particularly in comparison with more resource-intensive methods for investigating response activity such as think-aloud studies or eye-tracking.

References

Ben-Yehudah, G., & Eshet-Alkalai, Y. (2018). The contribution of text-highlighting to comprehension: A comparison of print and digital reading. *Journal of Educational Multimedia and Hypermedia*, *27*(2), 153–178.

Butler, L. (2020). *GCSE Geography: Strategies to support students in tackling level marked questions*. OCR.

Crisp, V., & Ireland, J. (2022). *A structure for analysing features of digital assessments that may affect the constructs assessed*. Cambridge University Press & Assessment.

Drijvers, P. (2019). Digital assessment of mathematics: Opportunities, issues and criteria. *Mesure et Évaluation en Éducation*, *41*(1), 41–66.

Goodwin, A. P., Cho, S.-J., Reynolds, D., Brady, K., & Salas, J. (2020). Digital versus paper reading processes and links to comprehension for middle school students. *American Educational Research Journal*, *57*(4), 1837–1867.

Hughes, S., Custodio, I., Sweiry, E., & Clesham, R. (2011, November 8–10). *Beyond multiple choice: Do e-assessment and mathematics add up?* [Paper presentation]. AEA-Europe 12th Annual Conference, Belfast, Northern Ireland, UK.

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *The Journal of Technology, Learning, and Assessment*, *4*(5).

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education*, *21*(3), 207–226.

Lemmo, A. (2023). Tasks in paper and digital environments: An exploratory qualitative study. *International Journal of Mathematical Education in Science and Technology*.

Lowrie, T., & Logan, T. (2015). The role of test-mode effect: Implications for assessment practices and item design. *Proceedings of the 7th ICMI-East Asia Regional Conference on Mathematics Education*, 649–656.

Ofqual. (2020). *Online and on-screen assessment in high stakes, sessional qualifications*. Ofqual/20/6723/1.

Oxford Science Team. (2019, November 22). Insights from the 2019 AQA GCSE Combined Science Trilogy exams. *Oxford Education Blog*.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 279–293.

Williamson, J. (2023). *The feasibility of on-screen mocks in maths and science*. Cambridge University Press & Assessment.

Appendix

Table A1 shows the number of item attempts and annotation rates for all items in the study. Items that appeared on both Foundation and Higher tier papers share an item reference (e.g., B03-F and B03-H for the Foundation and Higher tier instances of the same Biology item).

Table A1: Annotation rates by item

Item	Tier	Label	Description	N attempts	Annotation rate	
					Overall	Attempts only
B01	F	1	MCQ pick correct term	999	0.09	0.09
B02	F	2	MCQ with diagram	999	0.13	0.13
B03-F	F	9	MCQ with graph	993	0.25	0.25
B03-H	H	1	MCQ with graph	997	0.33	0.33
B04-F	F	10	MCQ with table	998	0.16	0.16
B04-H	H	2	MCQ with table	997	0.31	0.31
B05	H	9	MCQ parallel statements	1000	0.21	0.21
B06	H	10	MCQ calculation	997	0.69	0.69
B08	H	15b	Deduce using graph	987	0.44	0.44
B09-F	F	17	Multi-part algae question	964	0.54	0.54
B09-H	H	11	Multi-part algae question	1000	0.69	0.69
C01	F	1	MCQ pick correct term	997	0.08	0.08
C02	F	2	MCQ chemical equation	998	0.22	0.22
C03	F	9	MCQ calculation	994	0.40	0.40
C04-F	F	10	MCQ parallel statements	999	0.13	0.13
C04-H	H	1	MCQ parallel statements	1000	0.19	0.19
C05	H	2	MCQ pick correct term	999	0.13	0.13
C06	H	9	MCQ shell diagram	999	0.49	0.49
C07	H	10	MCQ calculation	996	0.74	0.74
C08	F	14a	State empirical formula	845	0.31	0.36
C09-F	F	16d	Explain term “oxidised”	884	0.38	0.42
C09-H	H	11d	Explain term “oxidised”	997	0.51	0.51
C10	H	13d	Calculate moles	958	0.62	0.64
P01	F	1	MCQ 4 parallel diagrams	997	0.15	0.15
P02	F	2	MCQ definition	996	0.10	0.10
P03-F	F	9	MCQ with table	995	0.13	0.13
P03-H	H	1	MCQ with table	999	0.22	0.22
P04-F	F	10	MCQ with diagram	996	0.54	0.54
P04-H	H	2	MCQ with diagram	1000	0.42	0.42
P05	H	9	MCQ calculation	1000	0.52	0.52
P06	H	10	MCQ calc with diagram	998	0.53	0.53
P07	F	15ci	Trolley acceleration	857	0.42	0.44

Item	Tier	Label	Description	N attempts	Annotation rate	
					Overall	Attempts only
P08	H	15b	Trolley acceleration	913	0.28	0.30
P09	F	13a	Calculate force	960	0.85	0.85
P10	H	13bi	Fleming's rule	976	0.27	0.27
M01	F	7	Angle problem	949	0.50	0.52
M02-F	F	18	Word problem	940	0.37	0.39
M02-H	H	7	Word problem	1000	0.42	0.42
M03-F	F	15b	Short word problem	922	0.35	0.38
M03-H	H	3b	Short word problem	995	0.56	0.56
M04	H	16	Angle problem	955	0.88	0.91
M05	F	1a	Write name of polygon	873	0.23	0.25
M06	F	4ai	Next term in sequence	996	0.49	0.49
M07	F	7	Partial Venn diagram	996	0.89	0.89
M08	H	7	Work out perimeter	986	0.93	0.94
M09	H	12a	Next term in sequence	998	0.56	0.56
M10	H	13	Algebraic graph	960	0.34	0.35

Learners' annotations and written markings when taking a digital multiple-choice test: What support is needed?

Victoria Crisp (Research Division), **Sylvia Vitello** (Research Division), **Abdullah Ali Khan** (Education Digital Products and Services), **Heather Mahy** (Education Digital Products and Services) and **Sarah Hughes** (Education Futures Directorate)

Introduction

While digital tests and exams are prevalent in many contexts, so far they are uncommon for general qualifications in England and for general qualifications available internationally that are based on the English assessment model. However, moves towards providing digital exams for appropriate general qualification contexts are now progressing at pace. Digital exams offer a variety of potential advantages over paper-based exams, for example: providing a better match to how learners conduct their school work in some subjects; assessing skills that are difficult to assess authentically in a paper-based exam (e.g., computer programming); providing customisable accessibility features that candidates can adjust to their needs (e.g., screen background colour); allowing each candidate to individually control the playback of an audio or video stimulus; and making it easier for candidates to edit their responses. Nonetheless, care is needed to ensure quality and fairness as digital exams are introduced. One of many factors that needs to be considered is ensuring that candidates are not hindered in how they work during the exam and that the digital testing platform, its functionality, and any accompanying support materials (e.g., scrap paper) allow candidates to use their relevant knowledge, understanding and skills to select or produce their answers. This relates to validity since it contributes to ensuring that candidates' results reflect relevant constructs and can be interpreted and used in the intended ways (Messick, 1989).

As well as ensuring validity, comparability also needs to be considered. In contexts where parallel digital and paper-based exams are offered as alternatives, careful thought needs to be given to the intentions for comparability between modes (see Shaw, Crisp & Hughes, 2020, for a framework to support thinking about the different kinds of claims that might be made regarding comparability between assessments). One important aspect of this is the extent to which cognitive

processes (as supported by tools and materials) differ between modes and whether any differences are appropriate given the comparability claims that are made.

These considerations around validity and comparability have led to discussion within Cambridge University Press & Assessment about the annotations that candidates make on their exam papers in paper-based exams, the functions that these serve in terms of supporting candidates' question-answering processes and how such functions can be appropriately supported when candidates take digital exams. As a starting point, Williamson (2025, this issue) analysed candidates' annotations in paper-based GCSE Maths and Science exam scripts. She found that annotations occurred quite frequently (overall rate of 40 per cent across all questions considered) and that rates of annotation varied for different questions (from 8 to 93 per cent). The types of annotations observed to have been made by candidates on their paper-based exams included highlighting key information, crossing or ticking response options in multiple-choice questions (MCQs), annotating the question with related facts or rules, annotating a graph or figure, and showing working out.

Several other studies have explored how learners use scrap paper to support their exam techniques when taking digital tests, often with either the same learners or parallel groups of learners attempting digital and paper-based versions of the same or similar tests in order that comparisons can be made. One finding is that some learners transfer material (e.g., diagrams) from screen to scrap paper when taking a digital test so that they can annotate, and that this transfer can sometimes lead to errors (Johnson & Green, 2006; Hughes et al., 2011). Another important finding is that learners tend to write less rough work on scrap paper during digital tests than on the test paper during a paper-based test (Johnson & Green, 2006; Hughes et al., 2011; Pengelley et al., 2023; Nastuta & Liu, 2023). It has been theorised (using cognitive load theory) that this may be a result of an additional cognitive load cost that may be incurred when switching attention between modes (Pengelley et al., 2023).

Another possible contributor to reduced written working and markings on scrap paper during digital tests is simply that certain written actions are not possible, or not as easy or natural, on scrap paper compared to on a paper-based test. For example, to annotate a diagram provided in a question, as exemplified in Hughes et al. (2011), learners taking a digital test would have to copy the diagram to scrap paper before being able to add annotations. Settlage and Wollscheid (2024) explored how being able to write on a paper test might contribute to mode effects. They allowed one cohort of learners taking three paper-based multiple-choice tests to write on their test papers while another cohort (taking the same paper-based tests) were only allowed to write on scrap paper. After controlling for differences in ability, the findings showed that those who were allowed to write on their test paper performed significantly better on two out of three tests, and overall. Performance improved by 3.5 per cent overall. Settlage and Wollscheid (2024) do not report the volume of written work from each cohort, except to note that they were surprised that over 90 per cent of learners who were instructed not to write on the test papers wrote nothing on their scrap paper. Their findings

imply that mode switching is not the only factor affecting the extent to which learners make annotations and written markings on scrap paper during digital tests.

The current research was specifically motivated by Cambridge International Education's plans to introduce digital versions of IGCSE multiple-choice exams in economics, accounting and the sciences running in parallel to existing paper-based exams. As well as potential advantages for the efficiency of processing and marking responses, for candidates it removes the current need to record their responses on a machine-readable form separate to the exam paper. To support this development, this research set out to enhance our understanding of the exam techniques and types of written annotations or markings that learners may wish to be able to use to support their thinking when taking digital multiple-choice exams. It was hoped that the findings would provide insights to inform any necessary additional developments to testing platform functionality and inform decisions about the need for any accompanying materials (e.g., scrap paper). Additionally, the research aimed to further explore issues around the factors that contribute to learners writing less rough work and markings on scrap paper during a digital test than they write on paper-based tests. To explore these themes, we asked learners to take a digital multiple-choice test (based on IGCSE Economics questions) while having access to either scrap paper or a print of the test. The inclusion of a test print was also interesting from the perspective that this could potentially be an option that testing organisations could consider providing to learners early in the introduction of digital exams, or to learners with certain kinds of learning needs. That said, a test print may not be an elegant solution to supporting learners' cognitive processes during a digital exam, given it involves duplication of material.

Since written (or sketched) work on scrap paper cannot technically be considered annotation (unless a learner reproduces part of the question or stimulus first), in this article we refer to both "annotation" and "written markings" to mean any writing (or sketching) on the test print or on scrap paper.

Method

Participants

The participants were 52 learners in three schools in England, aged around 17 years old, who were studying A Level Economics. Unfortunately, it was not possible to involve learners studying IGCSE Economics due to practicalities around timing and access. As a result, the test questions are likely to have been slightly easy for the learners, which should be kept in mind when interpreting the findings.

Materials

Digital test

A digital multiple-choice economics test was prepared using 15 questions from a past IGCSE exam paper (designed for 16-year-olds). The questions included a range of common design features, such as a stimulus diagram, a stimulus table, calculation, text-only response options, and tabulated response options.

In addition, a short digital test containing three other multiple-choice economics questions was prepared for demonstration purposes.

Test print

A PDF of the test (downloaded from the testing platform) was printed, double-sided on A4 paper.

Scrap paper

Scrap paper booklets were created to include two A4 sheets of lined paper followed by two A4 sheets of plain paper.

Procedure

The procedure was as follows:

- **Introduction** — The research and what would be involved in participation was described to learners, and informed consent was gained.
- **Platform demonstration** — The testing platform was demonstrated to learners using the demonstration test. Learners were shown navigation and other available functionality (i.e., a notes tool, calculator, question flagging tool).
- **Paper materials assigned** — Each learner was assigned to one of two conditions by being given either a print of the test or a scrap paper booklet, at random. It was emphasised to learners that they could use this paper as much or as little as they liked.
- **Test** — Learners were given up to 25 minutes to attempt the digital test. They used their own devices (laptops; tablets with or without keyboards).
- **Observations** — During the test, some learners were observed by a researcher, with notes on each learner's interactions with screen and paper captured in an observation sheet.
- **Questionnaire** — After the test, all learners completed an online questionnaire. This asked about their views and experiences of the digital test, their exam techniques and ways of working, and their use of the paper materials.
- **Interviews** — After the test and questionnaire, most learners were interviewed. Each interview usually involved a pair of learners where each learner had been given a different type of paper material. The interviews were semi-structured and explored learners' views and experiences in more depth.

The numbers of learners who participated in each element of the research are shown in Table 1. Learners were assigned random IDs, in the form "L01".

Table 1: Number of learners taking part in each research task

Research task	Number of learners		
	Test print	Scrap paper	Total
Digital test	27	25	52
Observation during test[1]	15	9	24
Post-test questionnaire	27	25	52
Post-test interview	22	23	45

1 The imbalance in the numbers of learners in each condition who were observed arose because of practicalities in the classrooms during data collection sessions.

Ethical considerations

The research protocol was reviewed by Cambridge University Press & Assessment's Research Ethics Committee and received a favourable outcome. Only learners who gave informed consent participated in the research.

Analysis

Initial analysis of observation notes involved counting the number of observed learners who interacted with the paper materials in any way. Then, further analyses explored how frequently learners used the paper when attempting questions, how much learners switched between modes, and which mode they spent more time on. Many of the observed learners carried out a main round of attempting the test questions, followed by a round of checking their responses. The analyses focused on each learner's main round of attempting the questions.

The paper materials given to learners for use during the test were inspected for whether there was any written work, drawing or annotations and, if so, how much space was used. Additionally, the kinds of annotations and written markings made by learners were categorised.

To analyse the questionnaire data, frequencies and percentages of each response were obtained. The interview transcripts were analysed using thematic analysis.

Findings

The findings are organised by overarching theme, with evidence from the different data sources included as relevant. We begin with evidence on the amount of written work conducted by learners and the possible effect of the nature of the assessment on this. We then look at the types of annotations and written markings that learners reported using in paper-based tests, and that they used during the digital test taken for the research. Next, we consider learners' views on the roles served by annotations and written markings, on having access to paper during the digital test, and on the functions that paper materials can support. Finally, we cover how learners interacted with the paper materials, the impact of switching attention between modes, and the impact of mode on the amount of mental working.

Amount of written work

In total, 25 out of the 52 learners (48 per cent) made some written markings on the paper materials they were given. This same proportion was found in both paper conditions (13 out of 27 learners in the print condition and 12 out of 25 learners in the scrap condition wrote on the paper).

Learners' paper materials were inspected for markings relating to each test question. This revealed wide variation between learners in terms of how many questions they made annotations for (Figure 1). A small number of learners, most of whom received the test print, made annotations for all, or almost all, questions. Other learners made annotations for less than half of the questions. Among these learners, those with scrap paper tended to annotate for fewer questions than those with the test print.

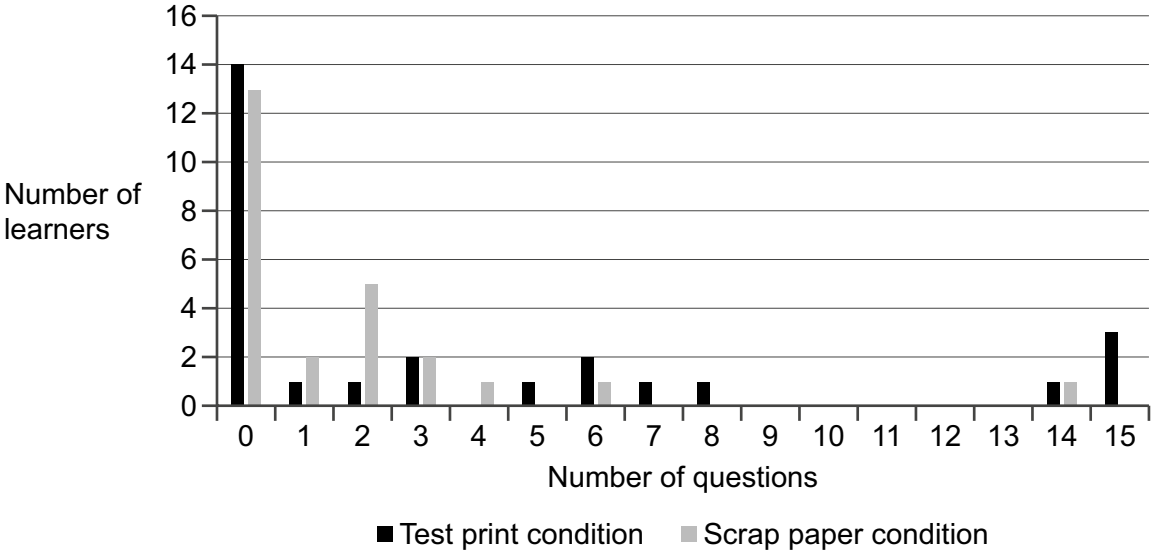


Figure 1: The number of questions for which learners made any written markings on paper

The analysis also considered how much physical space learners used on the paper materials. In the print condition, of the 13 learners who wrote on the paper, eight used moderate amounts of space in or around some questions, two used minimal amounts of space in or around one or two questions, and one used large amounts of space in or around most questions. The remaining two learners used blank space on the first page of the test print as if it were scrap paper (i.e., their written markings were not on or next to the relevant questions).

In the scrap paper condition, 10 of the 12 learners who wrote on paper used one side of the lined paper in the scrap paper booklets provided to them. Five of these learners used less than a third of a side, while the other five used more than 50 per cent of one side. Two learners wrote on two or three sides of lined paper.

None of the learners used the plain paper that was provided after the lined paper.[2]

Effect of assessment context on paper use

Some comments made during interviews suggested that learners might have used the paper materials differently if the test had been different in some respect. Learners' points are summarised in Table 2. Comments suggest that the paper materials may have been used more if the questions had been more difficult, involved more complex diagrams or calculations, or if the test had higher stakes.

2 From this evidence, we cannot tell whether learners with scrap paper would have used the plain paper rather than the lined paper if the plain paper sheets had been arranged first in the scrap paper booklets. The questionnaire asked the learners who had received scrap paper for their views on the type of paper they would prefer to have when taking a digital test. Most preferred lined paper (11 learners) or did not mind what kind of paper they received (seven learners). A few learners preferred plain paper (three learners) or reported that they would like both lined and plain paper (three learners). (One learner gave no response.)

Table 2: Learners' interview comments regarding the effect of the nature of the test or questions on their use of paper materials (summarised comments)

Test print	Scrap paper
• Did not need the print for this test but useful for other tests. • Likely to use the test print more for more difficult questions. • Might have looked at the test print if there had been a more complicated graph to analyse. • Did not need the print for this test but it would be useful for a big graph or a table with lots of data. • Might draw diagrams on the print for A Level MCQs. • For A Level long-answer questions they might use the test print to annotate the question or draw the structure of their point of view or evaluation system.	• More likely to use scrap paper for more complex or more difficult questions. • Might use scrap paper if planning a longer answer/essay. • Might use scrap paper for more difficult calculation questions. • Scrap paper is useful when they need to draw a diagram or graph. • Might have drawn diagrams on the scrap paper if the test had been high stakes. • Considered using the scrap paper for a question involving an equation or calculation, but then found it was easy to do without any written work.

Types of annotations and written markings

In the questionnaire, learners were asked which exam techniques, from a list provided, they usually use when taking paper-based tests (Table 3). Ruling out incorrect responses was most commonly selected, followed by sketching a diagram, graph or table, underlining or circling key words, annotating an existing diagram, graph or table, showing working out, and keeping track of questions to revisit later.

Table 3: Usual exam techniques used in paper-based tests (closed response)

What exam techniques do you usually use in paper tests, including when answering multiple-choice questions? Select all that you use.	**Number and percentage of learners who selected each option**	
I underline/circle key words	34	65.4%
I write down facts, formulae or theories related to the question	29	55.8%
I make notes on ideas	18	34.6%
I show my working out	32	61.5%
I rule out answers (A, B, C, D) as soon as I know they're wrong	46	88.5%
I annotate the graphs, diagrams or tables that are in the test	33	63.5%
I sketch my own graph, diagram or table	34	65.4%
I keep track of questions that I want to come back to	32	61.5%
I don't do anything in particular	1	1.9%
Other (please specify)	0	0.0%

The paper materials used by learners during the research provide insights into the kinds of written markings made by learners while attempting the digital test. Categories were developed to capture the types of markings that were apparent. Figure 2 shows the numbers of learners who made markings fitting each category.

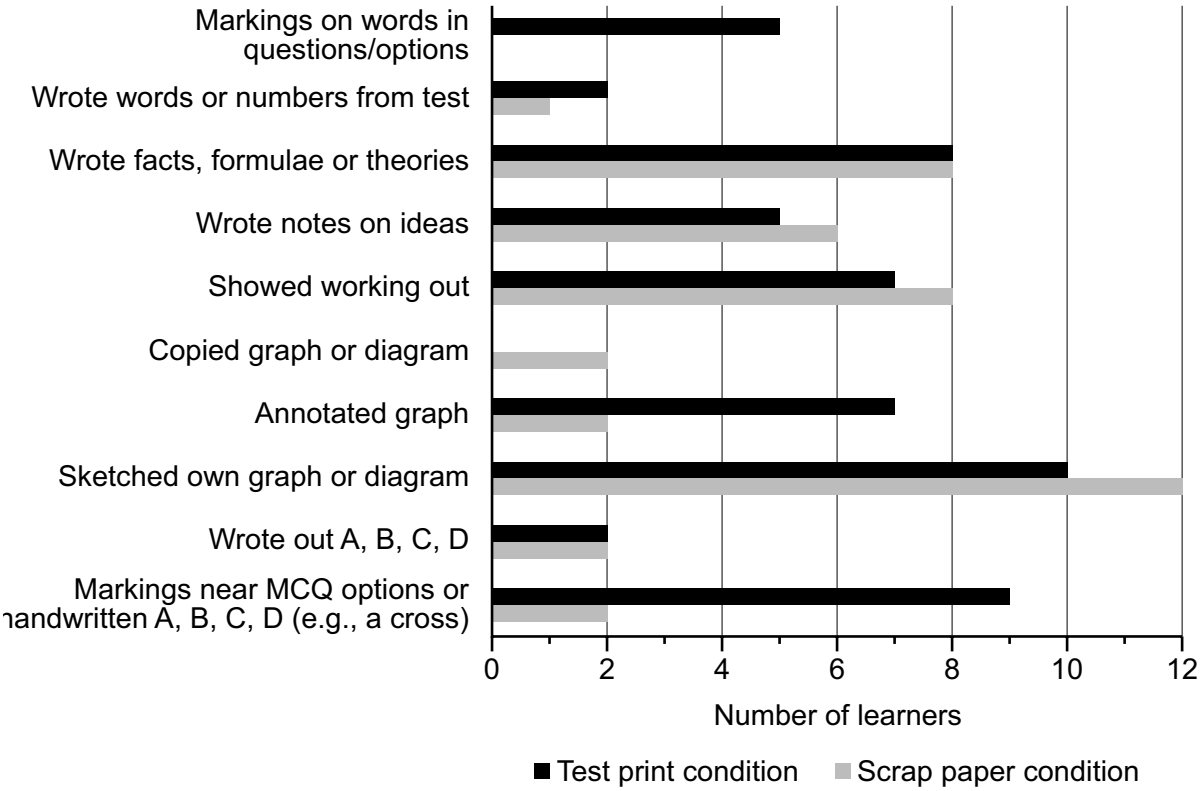


Figure 2: The number of learners in both paper conditions who made each type of annotation or marking on their paper. The categories are ordered thematically.

In both conditions it was relatively common for learners to sketch their own graph or diagram, show their working out, and to write down facts, formulae or theories. Other types of annotation or written markings were notably more common among learners in the test print condition than in the scrap paper condition. The most marked difference was found for markings on or near multiple-choice response options, presumably made to record ruling in or out a response option. Only two learners with scrap paper made any markings related to selecting response options (e.g., writing out the letters "A", "B", "C", "D" and then circling a letter or using a line to strike through some letters). Annotating a graph was also more common in the test print than on scrap paper, perhaps unsurprisingly since it was only possible on scrap paper if learners copied the graph first. Additionally, learners with the test print were more likely to identify key information (e.g., words, numbers) in the question or response options (by circling or underlining) than learners in the scrap paper condition (by writing down words).

For each category in Figure 2, the data showed a range of specific, and sometimes idiosyncratic, annotations or markings. To provide an illustration of this variation, Figure 3 and Figure 4 show some examples relating to response options, as made on the test print and on scrap paper.[3] As can be seen, learners used ticks, crosses, circles and strikethrough lines as part of their process of ruling in or out options.

3 See Williamson (2025, this issue) for illustrations of other types of annotation in paper-based exams.

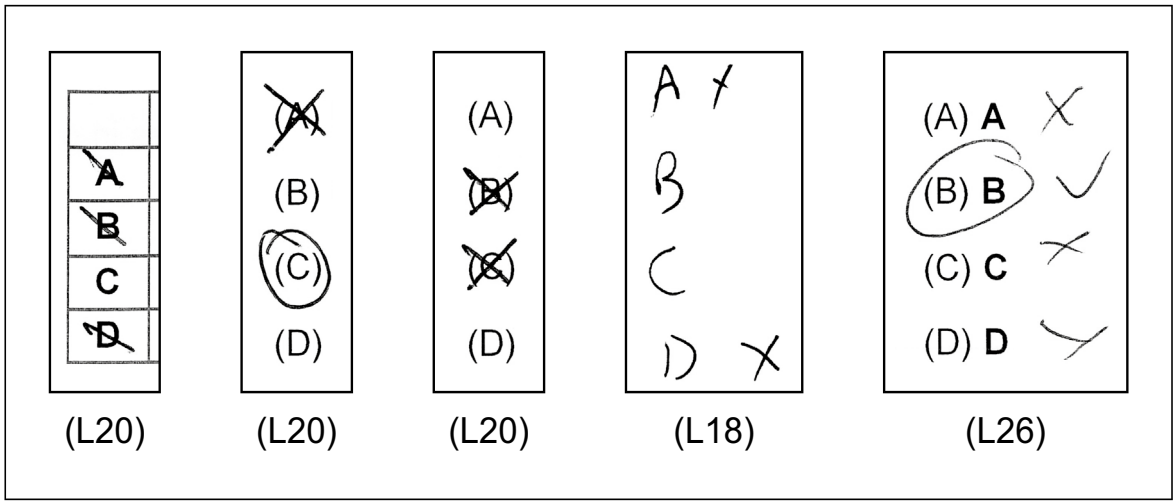


Figure 3: Examples of annotations and written markings relating to multiple-choice response options made in the test print[4]

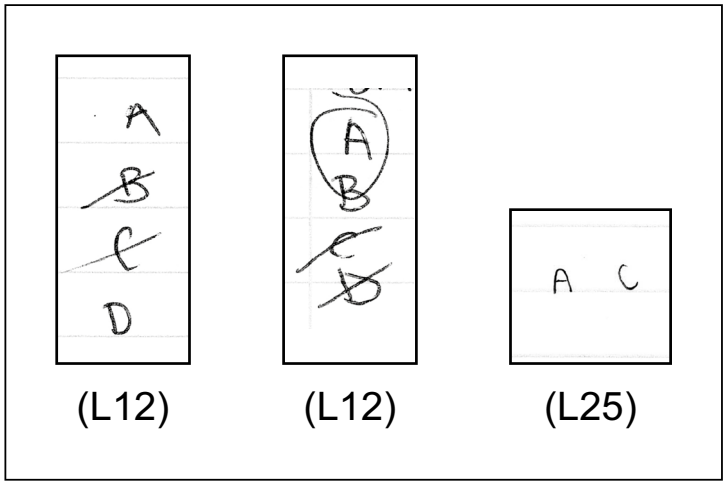


Figure 4: Examples of annotations and written markings relating to multiple-choice response options made on scrap paper

To summarise the variety of different kinds of written markings that were observed, Table 4 groups the categories from Figure 2 into five broader themes and lists illustrative examples. There was variation in the number of questions for which an individual learner used a particular type of annotation or written marking across questions. In some cases frequency of use was influenced by whether an annotation or marking type was appropriate to all or only to some questions (e.g., showing working out was only relevant to questions that involved calculation or a formula).

4 Note that learner L18 made their written markings on the first page of the test print.

Table 4: Annotation and marking types organised by theme, with exemplification

Theme	Examples
Extracting information from the question	• Underlining and circling key words. • Writing down key words from question context. • Writing down numbers required for calculations. • Copying information (e.g., percentages) from a table (without reproducing the table format).
Writing notes on subject content and ideas	• Writing down recalled economics formulae. • Writing down economic theory acronyms. • Writing down economics study mnemonics. • Using symbols to present information (e.g., up/down arrows). • Writing thoughts about the question or answer in short phrases or sentences.
Showing working out	• Arithmetic calculation steps. • Working out an economics formula numerically (e.g., inputting numbers). • Working out an economics formula conceptually (e.g., writing comments alongside it).
Graphs and diagrams	• Adding points (e.g., dots or circles) to a graph in the test question. • Adding arrows to a graph in the test question. • Shading areas of a graph in the test question. • Sketching one or more versions of a new graph. • Sketching a number line.
Response options	• Writing down letters "A" "B" "C" "D". • Putting a cross on (or next to) response options, presumably to rule them out. • Striking through options with a line, presumably to rule them out. • Circling one response option, presumably to indicate the correct response. • Putting a star next to one option, presumably to indicate the correct response. • Putting markings in tables next to response options (either to rule out options or mark-up plausible options). • Putting a "?" next to options.

Roles served by annotations and written markings

The interviews provide evidence regarding learners' views on the roles played by written markings in their question answering. One key theme was that annotations or written markings could support learners' thinking or working out and allow them to see their whole chain of reasoning. For example:

> "I mean for me I find it easier to kind of just write down my chain of thinking and reasoning and just having that in front of me … when it is in front of me, I can visualise it, it is a bit easier to kind of translate that into how I would work out the answer. So I feel that, for me, it's quite important, especially when the questions are a bit more complex." (L25, scrap paper)
>
> "I'm the kind of person that needs to write down all my thoughts." (L46, test print)

Two learners (interviewed together) felt that making notes could help unpack the meaning of key terms:

> "I would find a word and then I had to figure out like the key term. 'Oh, 'inelastic' blah blah blah' and like figure that out and write it down and then follow with my train of thought." (L12, scrap paper)

For another learner, their usual practice of writing notes helped identify which of their ideas were relevant:

> "I had nothing to scribble down any ideas. Because, even if I don't necessarily use what I scribble down, it kind of helps me filter through information to see what is relevant to the question." (L24, test print)

A common theme was that being able to annotate a stimulus graph or diagram, or to sketch a stimulus graph or diagram, supported their thinking, for example by allowing them to visualise concepts:

> "For one of the questions, I did use a really simple diagram that I drew to sort of just help me visualise what I think was the most prominent sort of choice." (L23, scrap paper)

> "Normally if I have a question with the graph, I'll try to annotate the graph ... it was a bit difficult to try and understand like, understand my answer without being able to draw on it." (L38, scrap paper)

Some learners commented on how annotations played a role in ruling in or out response options when answering multiple-choice questions:

> "I can't cross out the wrong answer on the computer and I also can't show my workings on the computer, so it's going to make my chain of analysis like a bit messy because I have to do everything in my brain." (L45, test print)

> "I feel better about myself if I know that I've only got two options left at one point ... when I'm able to cross my final one, I can like really focus on just my last option to make sure that I've got it right." (L24, test print)

Views on having paper materials during the digital test

In the questionnaire, learners who reported that they had not used the paper materials were asked whether they liked having the paper or would be happy not to have it. Eight of the 14 learners who did not use their test print, and 12 of the 13 learners who did not use their scrap paper, recorded that they liked having it. The remaining seven learners who had not used their scrap paper or test print reported that they would be happy not to have the materials they received.

Additional evidence comes from the interviews, during which learners were asked how they felt about having the paper materials they received. For each paper type, some learners reported that they liked having the paper materials and having them was comforting or reduced stress:

> "A sense of comfort because I know, if I ever need it, it's there for me to use." (L36, scrap paper)

> "I like having them, it was comforting." (L19, test print)

One learner felt that not having scrap paper would have affected their performance (presumably in a negative way):

> "I think, if I didn't have it, it would have – it would have affected my, like, answers." (L52, scrap paper)

Small numbers of learners in each condition felt that they should use the paper because they had been given it:

> "I kind of felt like I needed to use it, so I did. ... Because it's like, I don't know, it's just like there." (L27, test print)

Views on the functions that paper materials can serve during a digital test

During interviews, learners commented on the perceived benefits of the two types of paper materials, which often related to their understandings of how paper materials can support different types of written working and annotation. Table 5 summarises these comments. Scrap paper reportedly served various functions such as allowing learners to sketch diagrams and write down calculations or ideas, and supporting their thinking. For the test print, learners showed an awareness of the additional opportunities that this offered, such as annotating the text or a diagram, avoiding the need to re-draw diagrams, and making annotations near the relevant question.

Table 5: Learners' interview comments regarding benefits of the paper materials (summarised comments with some quotations for exemplification)

Test print	Scrap paper
• It can be annotated (e.g., underline key words, draw on diagrams, cross out options). • Aids quick annotation. • Annotations/notes can be made near the question so less confusing than scrap paper – you know which notes relate to which question. • More useful than digital notes tool which is separate from the text. • Reduces the need to switch between screen and paper, e.g., "I liked having it printed out because I could read it, and I didn't have to be going like back and forth between reading the question online" (L19, test print). • Avoids needing to re-draw diagrams, which would take up time. • Allows use of a pen to point at or focus attention on parts of the question. • Appropriate amount of space around questions to work. • Useful for checking/reviewing answers, e.g., "just an easier way to get my thoughts in check" (L26, test print). • Useful if there are multiple steps or if question involves numbers. • Easier to process the questions if you can write things down rather than doing it in your head. • Able to write down thoughts and do maths on paper.	• Useful for drawing diagrams/graphs (e.g., helps to visualise the answer). • Useful for calculations (e.g., writing down a multiple-step calculation). • Easier than doing working in head as they can see the work in front of them, which supports their thinking. • Can record train of thought / write down working out as they go. • Can write anywhere and get ideas down quickly. • Provides space for rough work. • More familiar with using scrap paper. • Would rather write on paper than type (e.g., using digital notes tool).

How learners interacted with paper materials during the digital test

The observations of individual learners allowed a more detailed exploration of the ways that learners interacted with the paper materials for each test question. In total 13 of the observed learners interacted with their paper materials in some way (nine of the 15 observed learners who had the test print, and four of the nine observed learners who had scrap paper). This includes learners who only looked at or touched the paper, as well as learners who wrote or made markings on the paper.

For the 13 observed learners who interacted with the paper materials in some way, the observation notes provide insights regarding:

- Level of interaction with paper – whether the learner interacted with the paper including annotating or making markings, interacted with the paper but only by looking at or touching it, or did not interact with the paper.
- Integration of the two modes – whether the learner started reading the question on screen or on paper (as an indication of whether their processing of the question was led by the screen or paper mode), and how many times the learner "visited" the non-leading mode from their lead mode (no visits, one visit, or multiple visits). Only visits that involved some processing of the question were counted (i.e., visits to screen were not counted if learners only carried out administrative tasks such as clicking the platform's "next" button to advance to the next question).
- Relative time on the two modes – whether the learner seemed to have spent more time attending to the screen, more time attending to paper, or a similar amount of time on each mode. This was based on subjective judgements made by the researchers during observations.

Figure 5 shows these three aspects of interaction side by side for each relevant learner. The left panel shows considerable variation between learners in the number of questions for which they interacted with paper. When learners interacted with paper, in most cases they made some written markings.

There was also variation between learners in terms of how they used the two modes (middle panel). Only one learner used paper as their lead mode. They read all questions on their test print and did not make any visits to the screen, except to input their responses at the end of the test. The other 12 learners were all screen-led for all test questions, but they varied in how much they switched between screen and paper.

The right panel shows that learners differed in how they divided their time between modes, although there seems to be less variation than in other respects. Two learners spent more time on paper than on screen for all, or almost all, questions. Four learners spent more time on screen for all questions. The remaining learners spent more time on screen for most questions, but for some questions spent more time on paper or a similar amount of time on both modes.

The variation found between the observed learners does not seem to depend on paper condition. Learners with scrap paper have patterns that would not look out of place among the patterns found for the learners with the test print.

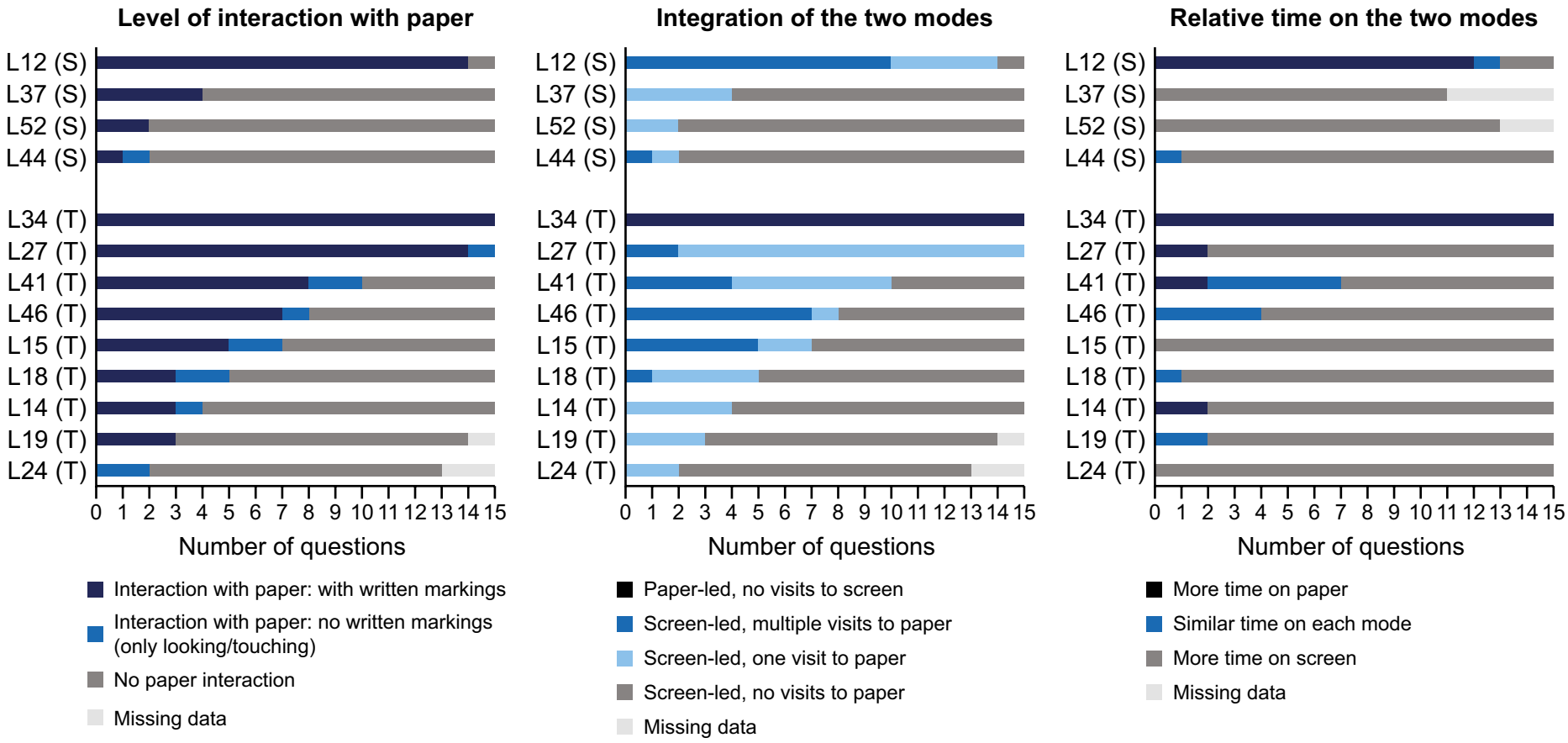


Figure 5: Profiles of interaction with paper for each of the 13 observed learners who interacted with the paper in some way.
S = Scrap paper condition; T = Test print condition.

To provide some relevant evidence about interactions with paper for all learners (i.e., not just those whom we observed), the questionnaire asked learners who received the test print whether they tended to read the test questions on screen or on paper first. Three learners stated that they usually read the questions on paper first, one responded that they read some questions on screen first and some on paper first, and 22 reported that they usually read questions on screen first.[5] This echoes the observation findings in the middle panel of Figure 5, and again suggests that using the test print as if it were a paper-based test was relatively uncommon.

Impact of switching attention between digital and paper modes

While analysis of the observations indicated the extent to which learners moved their attention between modes (Figure 5), learner comments during interviews provide insights into their experiences and views regarding mode-switching. Several comments suggested that switching modes was slightly inconvenient or confusing:

> "I had to continually switch from like scrap paper, question paper to iPad which was kind of inconvenient." (L46, test print)

> "It's a bit different having scrap paper and the test online because I have to look at the computer and look at my paper, it's just a tiny bit confusing but it's fine, yeah." (L05, scrap paper)

A related point was that switching modes appeared to act as a barrier to writing on the paper materials for some learners and that making notes or sketches would have been easier and more likely if this could be done in close proximity to the question:

> "If the test was on paper, like totally, then I'd probably write my workings out a little more." (L18, test print)

> "I didn't really do any [written work]. I might- I might have done some if it was on paper because then it's just easier I guess when the question's right there. I might have just drawn like a diagram or what numbers I use for a calculation. But it wasn't changed drastically." (L07, scrap paper)

Two learners (interviewed together) noted that switching between paper and screen interrupted their thinking:

> "If you look away, you look back, and you're, like: 'What was I thinking? Where was I looking?'" (L31, test print)

Another point about switching raised by three learners was that when answering a paper-based test they would already have a pen in their hand and that this, in itself, might increase the likelihood of annotation compared to when taking a digital test:

> "If I'm like doing it on paper, I would have a pen in my hand. It would be normal to just write it down." (L50, scrap paper)

Impact of mode on mental working

Comments from learners about doing more of the work "in their head" were common in the interviews. This often related to performing working out such as calculations, or holding ideas in working memory:

> "Usually I tend to write stuff out more, but with the computer I was more inclined to work stuff out in my head." (L18, test print)

5 The response was missing for one learner.

> "I kind of had to think more in my head and, like, remember everything that was going on." (L22, scrap paper)

One learner's comments suggested that conducting more thinking without writing anything down was partly a result of their lack of familiarity with the test being on screen:

> "I try to do it all in my head if I'm on the computer. ... I think it's because I don't have the ... thing right in front of me, where I can jot notes down or jot techniques down that I could use. I think it's just the position of the screen is slightly alien to me." (L08, test print)

Some learners noted that they did most of their working mentally during the test but that this was their usual practice for multiple-choice tests:

> "I kind of just did it in my head and I think that is what I'd normally do with multiple-choice questions." (L51, test print)

For other learners, the calculation question in the test was sufficiently easy that conducting the calculation mentally was unproblematic:

> "Even when I did the maths question I was like, 'I might write something down here', but then I was like 'I can just do it in my head, it's easy enough.'" (L21, test print)

However, a few learners worried about making mistakes, which sometimes led them to using paper to avoid this risk:

> "I also had to write down my calculations because I make a lot of mistakes when I do things in my head." (L46, test print)

Relatedly, some learners felt that it would be more difficult to keep track of their thinking or ideas when there were greater demands on working memory:

> "If you didn't [have paper] then you kind of just have to think about it in your head which can make it harder I suppose to remember what you want to say." (L51, test print)

Discussion

This research explored learners' practices in terms of annotations and written markings made on paper when taking digital multiple-choice tests. These are important considerations for ensuring that learners can effectively show their relevant knowledge, understanding and skills through the answers they select, and for ensuring comparability where paper and digital versions of an exam are offered in parallel (depending on the exact intentions regarding comparability). The research found that learners used, or would have liked to use, a range of written marking types such as making notes, writing down working out, annotating or sketching a diagram, circling or underlining words, and ruling in or out multiple-choice response options. Two key recommendations about how to support candidates' question-answering processes during digital exams arise from the research. Firstly, scrap paper should be provided to all candidates when they take a digital multiple-choice exam in economics or another subject where some questions tend to involve calculation, use of formulae or sketching visuals to support answering, as scrap paper can serve various useful functions (including

allowing candidates to use some similar strategies to those they may use during paper-based exams). Providing scrap paper during digital exams may also be appropriate for a wider range of subjects and question types. Secondly, providing easy-to-use digital functionality to support annotation is particularly important for those functions that scrap paper does not serve well (e.g., annotating a graph). While the test print better facilitated some annotation types, the percentage of learners choosing to use it was no higher than for scrap paper and, as mentioned earlier, it would not be an ideal long-term approach given the duplication of material.

Two limitations relating to our research sample should be kept in mind when interpreting the findings. The participating learners were not due to take high-stakes digital exams so had not received relevant exam preparation. Additionally, the participants were Year 12 economics learners in England and not learners who were studying for IGCSE Economics. Thus, the questions were probably relatively easy for these learners, which could have affected their behaviours. Learners tended to report that their frequency of written markings might have been higher for more difficult questions, which could suggest that IGCSE learners would have annotated more frequently than our participants. However, we cannot be sure of this, particularly as some evidence suggests that learners use scrap paper less for more difficult questions than for easier questions (Pengelley et al., 2023). Despite these limitations, the findings provide a diverse set of evidence with no indications that the types of annotations or written markings would have been different for learners preparing to take IGCSE Economics.

Williamson (2025, this issue), and questionnaire responses from the current research, show that learners use various exam techniques in paper-based exams that involve making annotations or written markings (that are not part of their responses) on the exam paper. There was considerable variation in whether and how much our research participants used paper materials when taking the digital test. This may reflect exam-taking practices that learners have been taught or their own preferred ways of working during exams that have developed over time. It might also have been affected by differences in learner ability (and, therefore, how difficult the questions were for different learners), different levels of motivation and engagement for a low-stakes experimental test, and expectations around whether they should conduct more working mentally when taking a digital test. Additionally, a lack of familiarity and preparation for using paper alongside a digital test may have affected learners differently. When learners are preparing to take high-stakes digital exams, they will have opportunities to undertake practice tests and are likely to receive guidance around the ways that they can use digital platform functionality and any accompanying paper materials for annotations and written markings. Hopefully, such preparation opportunities should help each learner use the available tools and materials in the ways that best support them.

Our findings suggest that for some types of written markings, scrap paper generally worked well. Learners could sketch their own graphs or diagrams, note down ideas, facts, formulae or memory aids, and record working for calculations. Some learners did comment, however, that it was less confusing to write next to the relevant question in a paper-based test, but fundamentally scrap paper was

able to support these functions. For some other types of written markings, scrap paper did not work well. These included: marking a key word (e.g., by underlining or circling); annotating a graph provided in the test; and making markings to support ruling in or out multiple-choice response options. Small numbers of learners made efforts to perform these functions using scrap paper (e.g., copying a graph, writing out "A", "B", "C", "D"), but this was uncommon and viewed as inconvenient.

As discussed earlier, past research suggests that learners tend to write less working on scrap paper when taking a digital test than they do on a test paper when taking a paper-based test (e.g., Hughes et al., 2011; Johnson & Green, 2006). One possible explanation is that there is an additional cognitive load cost to switching attention between screen and paper, adding to working memory demands (Pengelley et al., 2023). Some learners' comments in the current research suggested that they felt they conducted more working mentally, and that mode switching was a contributing factor. The inclusion of the test print allowed us to consider whether past findings of learners writing less on scrap paper during digital tests (compared to the amount of rough work conducted in paper-based tests) may be partially a result of certain actions not being possible (or convenient) on this type of paper material. Our findings support this as a possible explanation. Learners receiving the test print were no more likely to write on the paper support materials overall than those who received scrap paper, but certain types of annotations or written markings were more common among those with the test print.

It also appears plausible that various other factors could contribute to reduced written work during digital tests, such as learners' expectations of how they should work during a digital exam and their existing test-taking habits. For example, some may expect not to use paper at all and have little past experience of using paper while taking a digital test. Additionally, there is a physical element as well as the cognitive element of switching; for example, some learners commented that in a paper-based exam they would already have a pen in their hand.

Taken together, it seems likely that several factors contribute to fewer written annotations and markings being made on scrap paper during digital tests than are made on paper-based tests: the cognitive load cost involved in switching mode; physical factors involved in switching mode; learner expectations and habits; and the functions that scrap paper can and cannot easily facilitate. In these ways, using scrap paper during a digital exam is different in nature to using the exam paper for annotation and rough work during a paper-based exam. The current research, and other prior research, exemplifies cases where learners chose not to write on paper even though they could have. The potential for the cognitive effort of switching attention between modes and other factors to act as barriers to paper use emphasises the need for teachers and exam providers to ensure that learners have opportunities to practise using scrap paper while taking digital tests so that this becomes familiar.

Conclusion

The current findings show that scrap paper can serve some of the important functions that annotations normally serve in paper-based exams. Therefore, we would argue that scrap paper should be given to all candidates taking a digital multiple-choice exam in economics or in another subject where some questions tend to involve calculation, use of formulae or sketching visuals to support answering. The provision of scrap paper during digital exams also seems likely to be appropriate across a wider range of subjects and question types, since some of the types of written markings observed in the current research are relevant to many exam contexts. For example, writing notes on subject content and ideas may be relevant to many question types, including essays, and for some learners this may be easier on paper than on screen. Additional research could add to our understanding of the importance of scrap paper for different contexts.

It is also important that digital testing platforms include appropriate and easy-to-use functionality to support annotations, particularly for those written marking types for which scrap paper does not work well. For the IGCSE Economics questions explored in the current research, these were: identifying key words, ruling in or out multiple-choice response options, and annotating a stimulus diagram or graph. Evidence from this research is feeding into the development of Cambridge University Press & Assessment's digital exams. For example, findings are informing prototyping and user experience testing of additional annotation functionality, which will then feed into platform developments. It is possible that, in time, the provision of appropriate digital functionality and increased learner familiarity with using such functionality could reduce the range of assessment contexts for which scrap paper needs to be given to all candidates.

Acknowledgement

With many thanks to various colleagues for their input during the planning of this research, to the learners who participated and to their teachers.

References

Hughes, S., Custodio, I., Sweiry, E., & Clesham, R. (2011, November 8–10). *Beyond multiple choice: Do e-assessment and mathematics add up?* [Paper presentation]. AEA-Europe 12th Annual Conference, Belfast, Northern Ireland, UK.

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *The Journal of Technology, Learning, and Assessment, 4*(5).

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (pp. 13–103). Macmillan.

Nastuta, S., & Liu, L. (2023, November 1–4). *TIMSS 2019 equivalence study: A quantitative approach to explore assessment mode effects on mathematics performance in England* [Paper presentation]. AEA-Europe 24th Annual Conference, Malta.

Pengelley, J., Whipp, P. R., & Rovis-Hermann, N. (2023). A testing load: Investigating test mode effects on test score, cognitive load and scratch paper use with secondary school students. *Educational Psychology Review, 35*(3), 67.

Settlage, D. M., & Wollscheid, J. R. (2024). Deconstructing the testing mode effect: Analyzing the difference between writing and no writing on the test. *Journal of the Scholarship of Teaching and Learning, 24*(2), 79–89.

Shaw, S. D., Crisp, V., & Hughes, S. (2020). A framework for describing comparability between alternative assessments. *Research Matters: A Cambridge Assessment publication, 29*, 17–22.

Williamson, J. (2025). How do candidates annotate items in paper-based maths and science exams? *Research Matters: A Cambridge University Press & Assessment publication, 39*, 66–89. https://doi.org/10.17863/CAM.116170

Research News

Lisa Bowett (Research Division)

The following reports and articles have been published since *Research Matters*, Issue 38:

Journal articles and other publications

Benton, T. (2025). Simultaneous linear equating for scenarios with optional test versions or across multiple alternative anchors. *Practical Assessment, Research, and Evaluation, 30*(1), 1.

Constantinou, F. (2024). 'If you have a question that doesn't work, then it's clearly going to upset candidates': What gives rise to errors in examination papers? *Oxford Review of Education*, 1–21.

Kreijkes, P., Kewenig, V., Kuvalja, M., Lee, M., Vitello, S., Hofman, J., Sellen, A., Rintel, S., Goldstein, D. G., Rothschild, D. M., Tankelevitch, L., & Oates, T. (2025, January 13). Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. Preprint.

Oates, T. (2025). England: Turbulent Years—PISA 2022 and COVID-19 School Disruption. In: N. Crato & H. A. Patrinos (Eds.), *Improving National Education Systems After COVID-19. Evaluating Education: Normative Systems and Institutional Practices*. Springer, Cham.

Oates, T. (2024, June). *Preparing for power: Policy making around the school curriculum from 2010*. Institute for Government [Guest Paper].

Research and statistics reports on our website

Gill, T. (2025). *The impact of reducing the number of exams on results in GCSEs.*

Gill, T. (2024). *Impact of taking Core Maths: Analysis of OCR specifications.*

Ireland, J., & Majewska, D. (2024). *How are learning theories used in national curriculum development?*

Miranda, A., & Gill, T. (2024). *Provision of GCE A Level subjects in 2023*. Statistics Report Series No. 140.

Miranda, A., & Gill, T. (2024). *Uptake of GCE A Level subjects 2023*. Statistics Report Series No. 141.

Rushton, N., & Majewska, D. (n.d.). *Timeline of mathematics education in England and the USA.*

Conference presentations

Kreijkes, P., Kuvalja, M., Kewenig, V., Hofman, J. M., Vitello, S., Lee, M., Sellen, A., Rintel, S., Tankelvitch, L., Goldstein, D. G., Rothschild, D., & Oates, T. (2024, October 16–17). *To bot or not(e): Effects of large language models and note-taking on text comprehension and retention* [Paper presentation]. Cambridge Generative AI in Education Conference 2024, Cambridge, UK.

The Association for Educational Assessment (AEA) annual conference took place in Paphos, Cyprus, on 6–9 November, https://2024.aea-europe.net. Our researchers presented four papers:

Constantinou, F. (2024, November 6–9). *What kind of contextualisation is appropriate for assessing application of knowledge? Towards a more comprehensive framework for embedding examination questions in context.*

Morley, F., Walland, E., & Vidal Rodeiro, C. L. (2024, November 6–9). *The performance of transformer-based auto-markers on science content: A scoping review.*

Rushton, N., & Crisp, V. (2024, November 6–9). *A new Comparative Judgement (CJ) approach: Exploring the potential of criteria-based CJ.*

Vidal Rodeiro, C. L., Gill, T., & Hughes, S. (2024, November 6–9). *Using assessment and response times data to evaluate a digital mock exams service.*

Blogs and podcasts

Benton, T. (November, 5). I reviewed all 866 of Ofqual's subject pairs visualisations (so you don't have to).

Greatorex, J. (October, 22). Cambridge at a global educational research event.

Lestari, S. (October, 18). Handwriting versus typing exam scripts: Evidence from the literature.

Lieber, E. (October, 8). Reflections on my first BERA: insights on justice, sustainability, and assessment.

Sharing our research

We aim to make our research as widely available as possible. Listed below are links to the places where you can find our research online:

Journal papers and book chapters

Research Matters (in full and as PDFs of individual articles)

Conference papers

Research reports

Data Bytes

Statistics reports

Blogs

Insights (a platform for sharing our views and research on the big education topics that impact assessment around the globe)

Our YouTube channel contains Research Bytes (short presentations and commentary based on recent conference presentations), our online live debates #CamEdLive, and podcasts.

You can also learn more about our recent activities from Facebook, Instagram, LinkedIn and X (formerly Twitter).

CAMBRIDGE

The Assessment Network

Assessment Horizons Conference 2025

Join us in Cambridge or online on 29-30 April for a two-day conference that will give you the opportunity to explore emerging and developing themes in assessment design, development and delivery.

For more details visit
cambridgeassessment.org.uk/events/assessment-horizons-2025

CAMBRIDGE

The Assessment Network

Postgraduate Advanced Certificate in Educational Studies: Educational Assessment

Transform your understanding of assessment with a postgraduate qualification from Cambridge

For more details visit

cambridgeassessment.org.uk/pgca

Contents / Issue 39 / Spring 2025

Cambridge University Press & Assessment
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

ResearchDivision@cambridge.org
www.cambridge.org