Research Matters / 39
A Cambridge University Press & Assessment publication

ISSN: 1755-6031

Journal homepage: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

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

Joanna Williamson

To cite this article: 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

To link this article: https://www.cambridgeassessment.org.uk/Images/research-matters-39-how-do-candidates-annotate-items-in-paper-based-maths-and-science-exams.pdf

Abstract:

Teachers, examiners and assessment experts know from experience that some candidates annotate exam questions. "Annotation" includes anything the candidate writes or draws outside of the designated response space, such as underlining, jotting, circling, sketching and calculating. Annotations are of interest because they may evidence aspects of candidates' response activity that would be overlooked when focusing on response spaces. We have some evidence on how candidates annotate their questions from mode effect studies comparing paper-based and digital assessments, but little information on which candidates annotate and how often they do so.

This article describes an exploratory study of annotations made by GCSE Combined Science and GCSE Mathematics candidates. The research analysed scripts from four random samples of 1000 candidates, one each from the Foundation and Higher tiers of each GCSE, and looked at the prevalence and types of annotation on different items. A particular motivation was to support the design of effective digital assessment in maths and science, through improving our understanding of candidates' response activity in these subjects.

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)

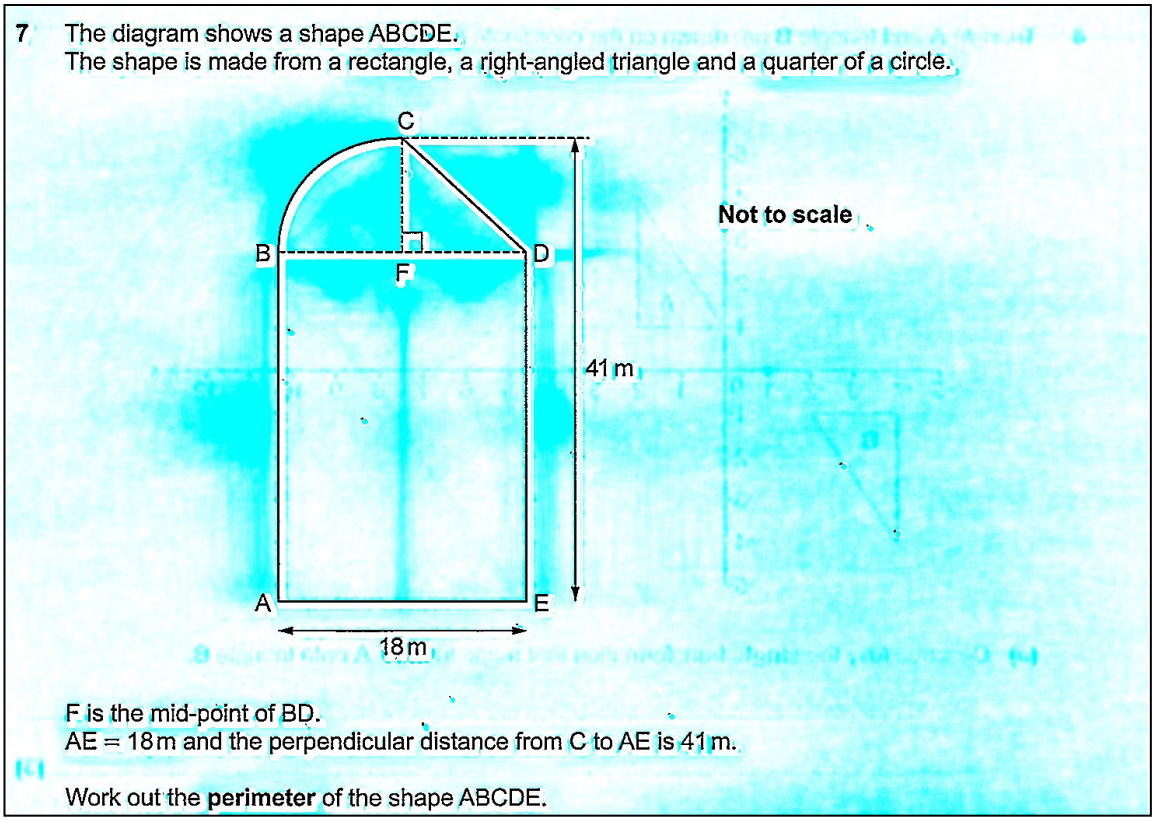

Figure 2: Heat map showing annotation on item M08 (Maths, Higher tier)

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.

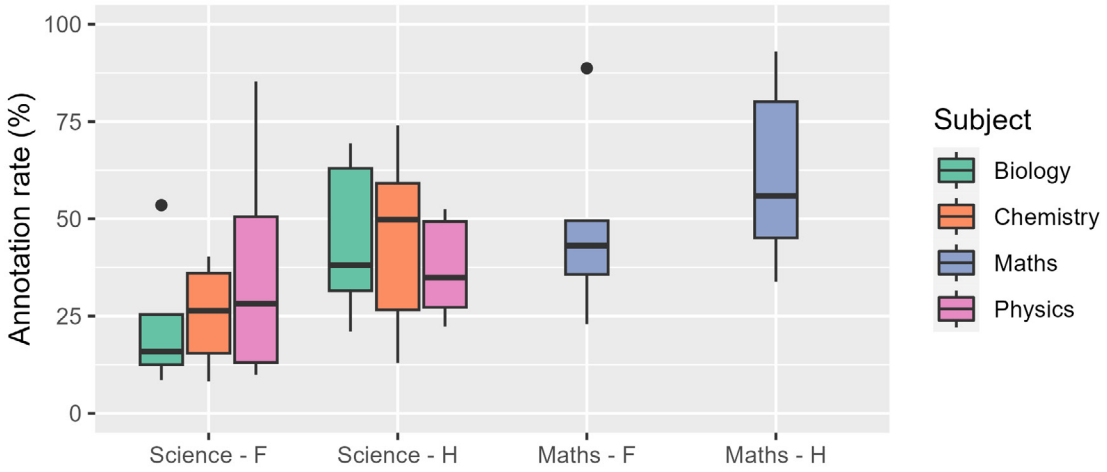

Figure 3: Annotation rates by GCSE, tier and subject area

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).

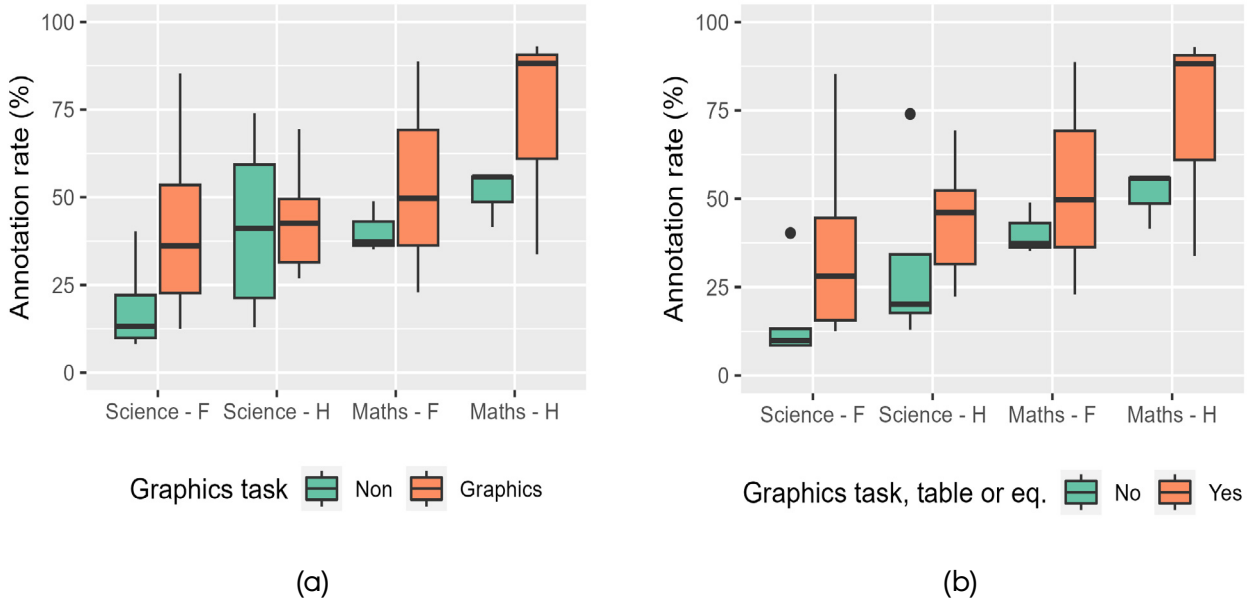

Figure 4: Annotation rates for graphics (including or excluding tables and equations) and non-graphics tasks

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.

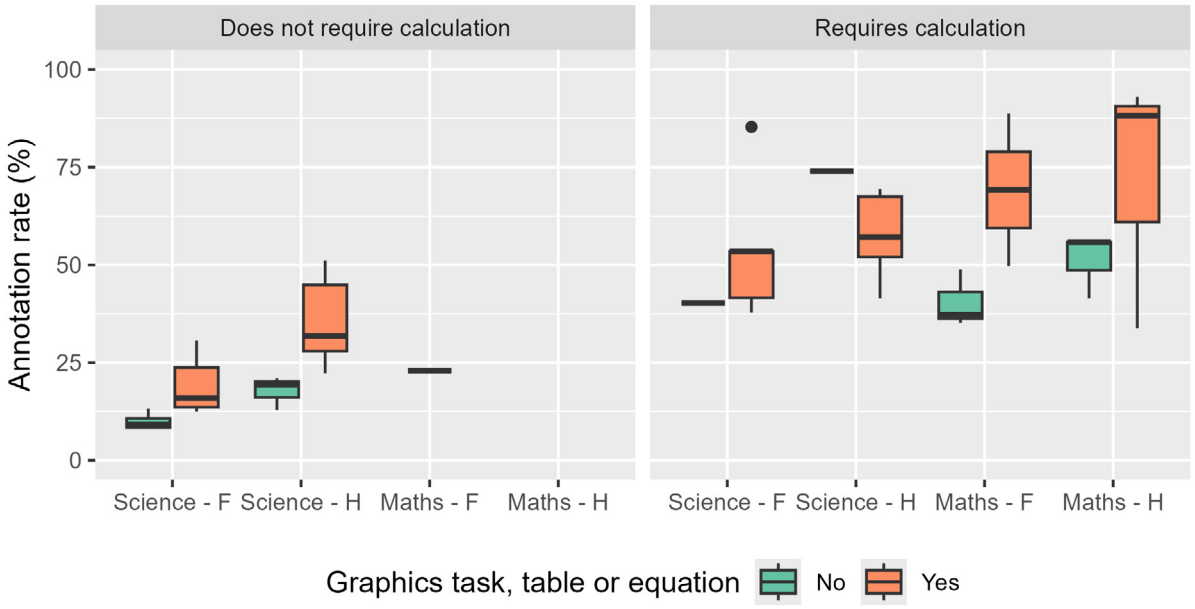

Figure 5: Annotation rates by graphics task (including items with a table or equation) and requirement for calculation

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.

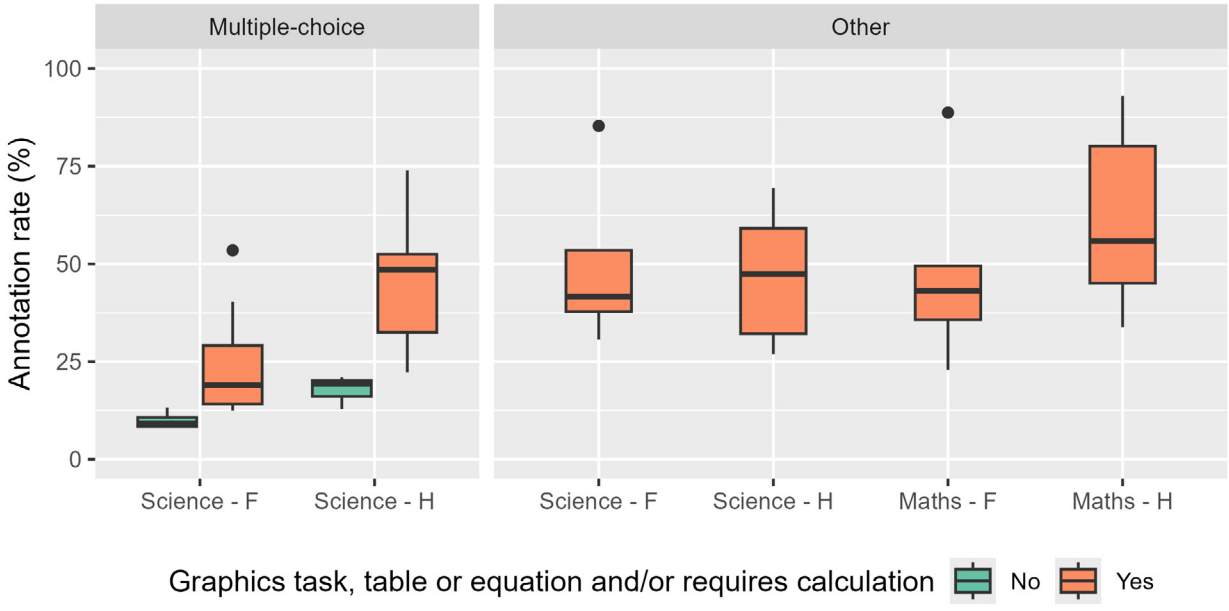

Figure 6: Annotation rates by question type (MCQ or non-MCQ), graphics task (including items with table or equation) and requirement for calculation

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).

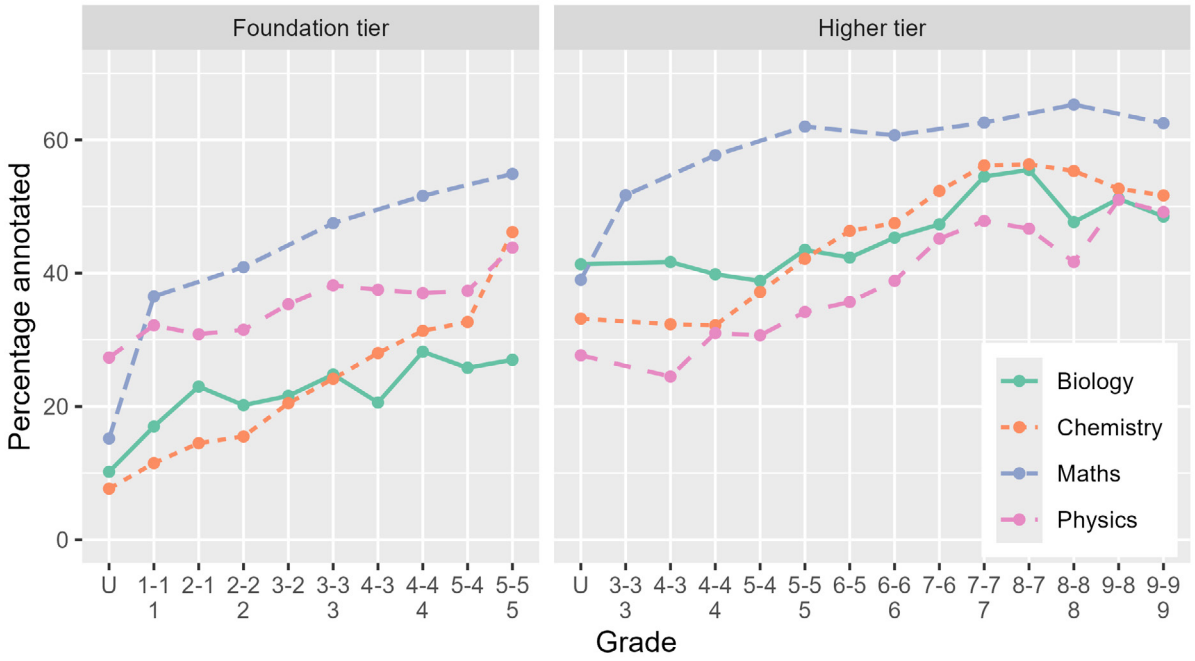

Figure 7: Percentage of item images annotated by relevant GCSE grade

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.

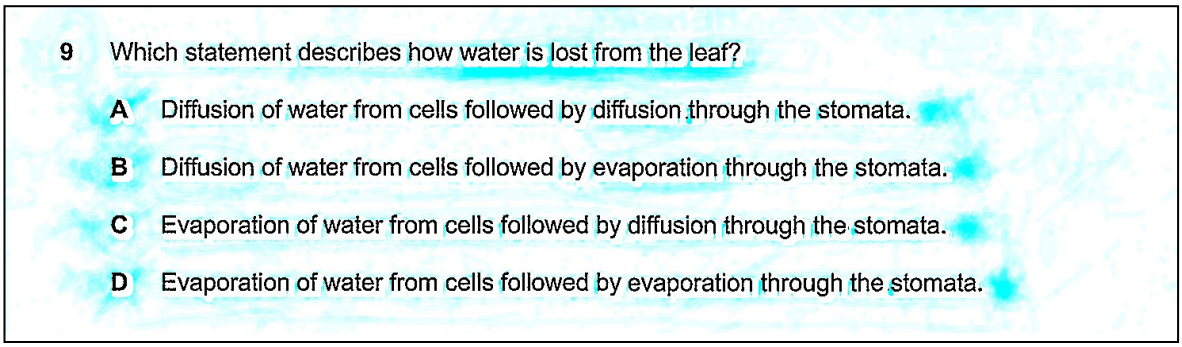

Figure 8: Heat map showing annotation on item B05 (Biology, Higher tier)

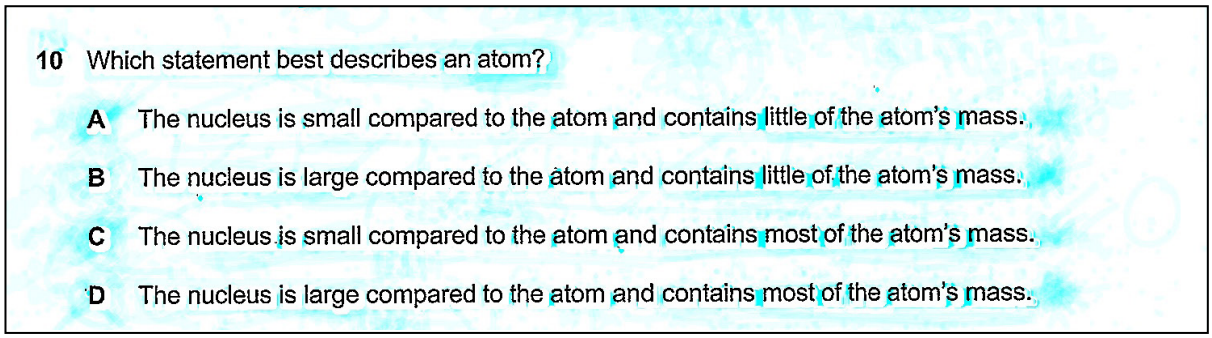

Figure 9: Heat map showing annotation on item C04-F (Chemistry, Foundation tier)

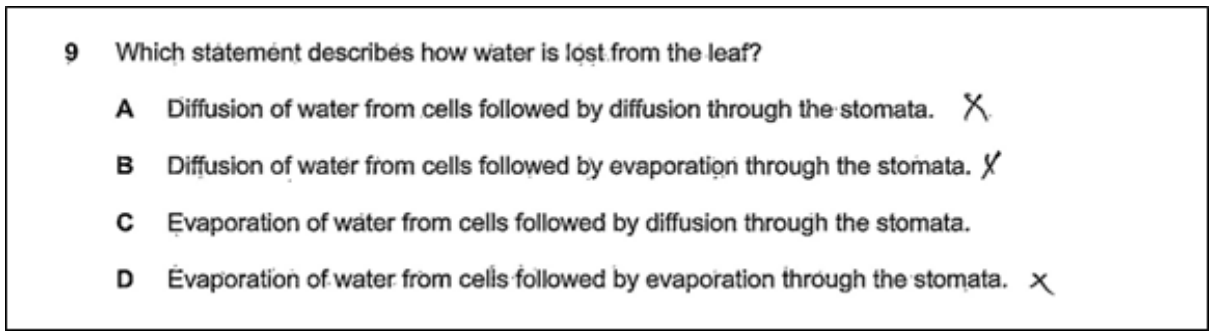

Figure 10: Example of candidate annotations marking crosses on three answer options, B05 (Biology, Higher tier)

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.

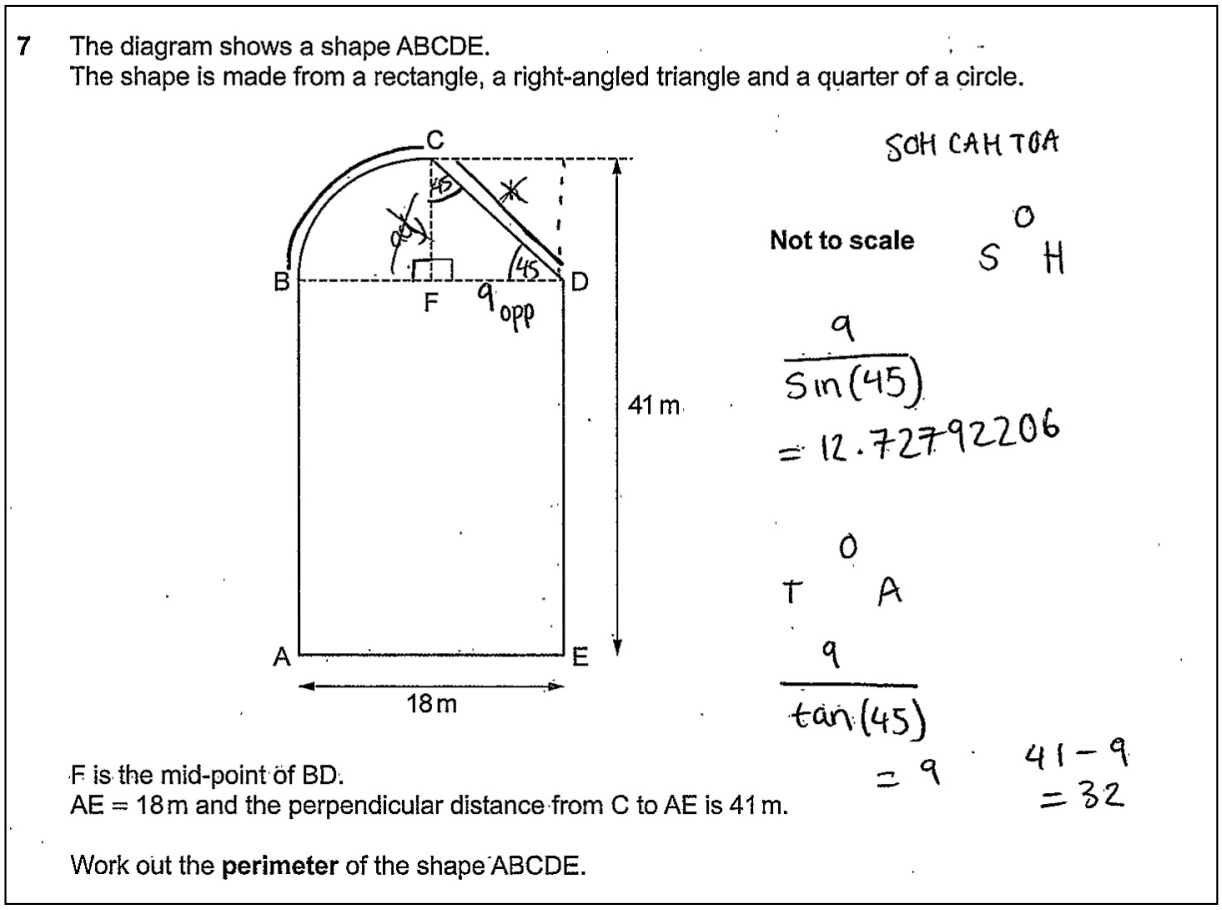

Figure 11: Example of candidate annotations noting angle facts, M08 (Maths, Higher tier)

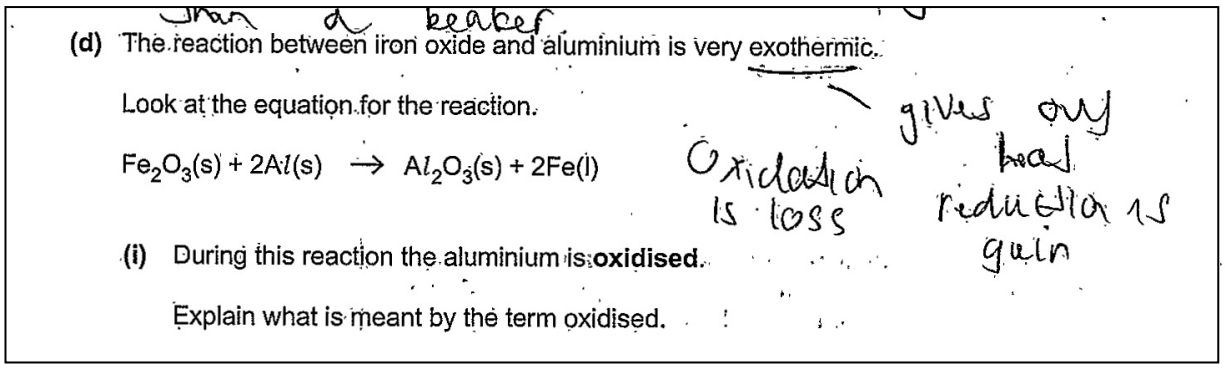

Figure 12: Example of candidate annotations noting information relating to the terms "exothermic" and "oxidised", C09 (Chemistry, Higher tier)

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.

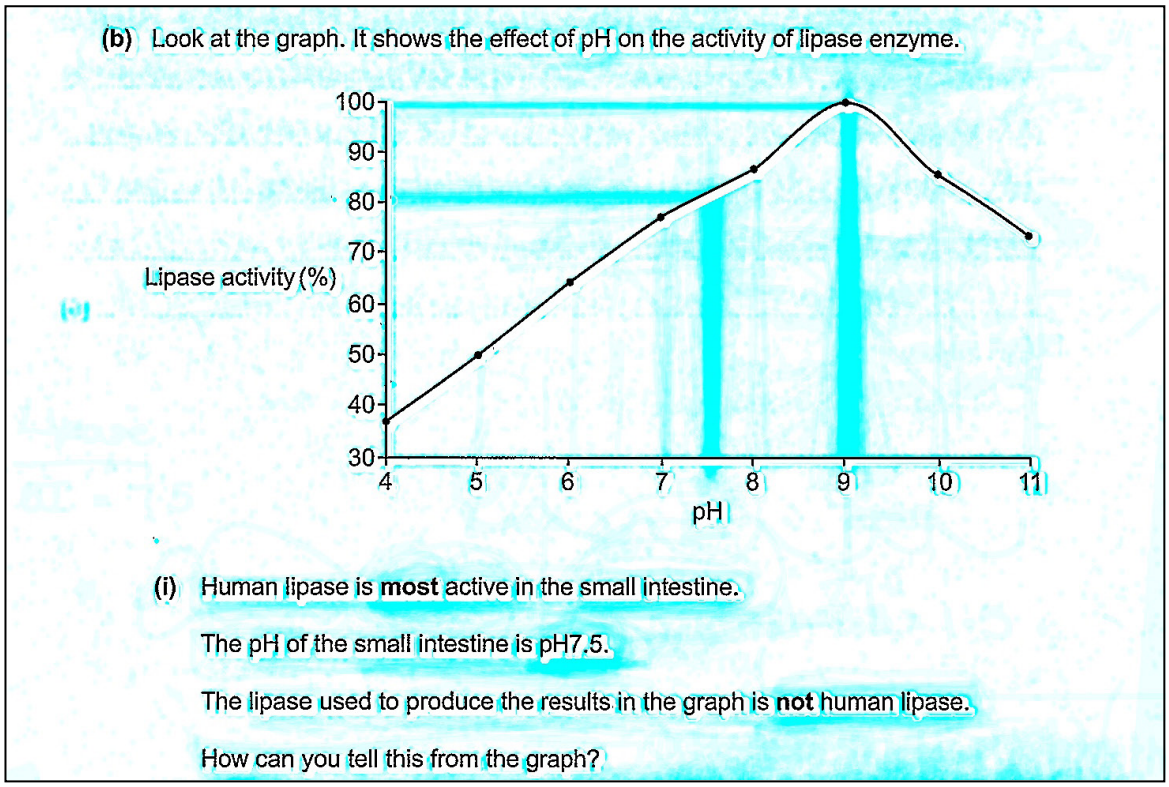

Figure 13: Heat map showing annotation on item B08 (Biology, Higher tier)

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.

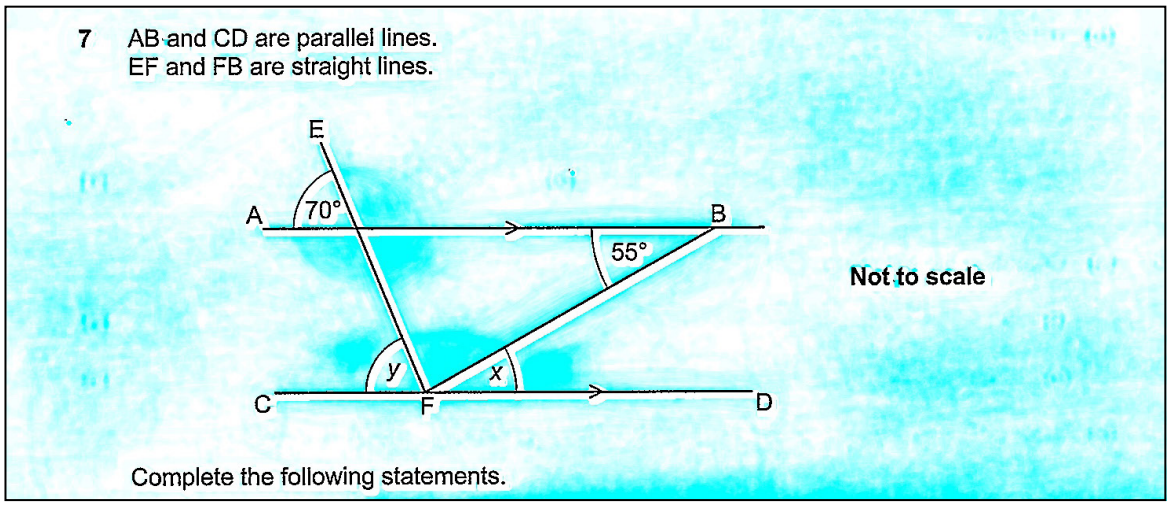

Figure 14: Heat map showing annotation on item M01 (Maths, Foundation tier)

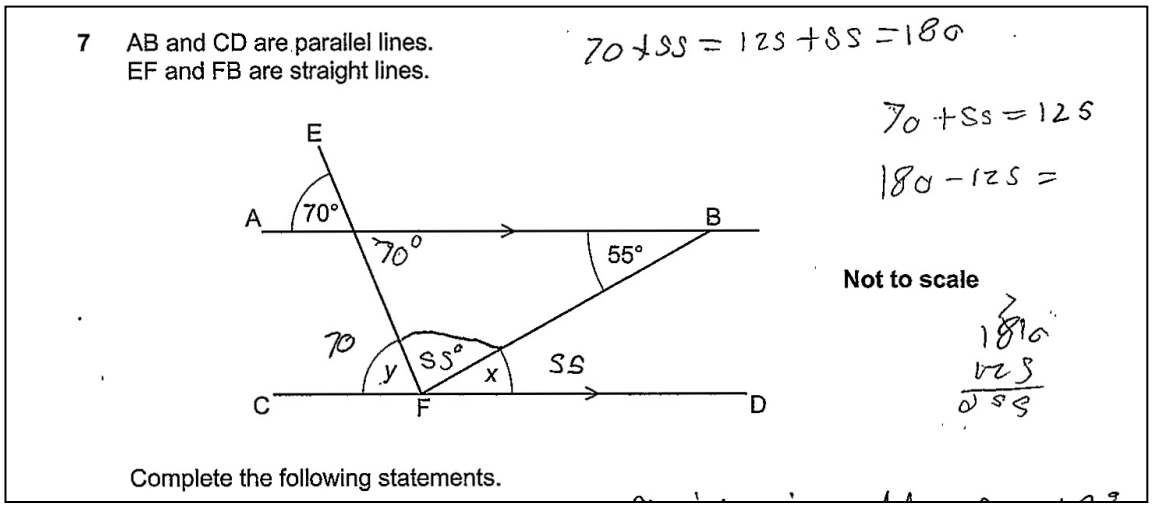

Figure 15: Example of candidate annotations suggesting use of several rules regarding angles, M01 (Maths, Foundation tier)

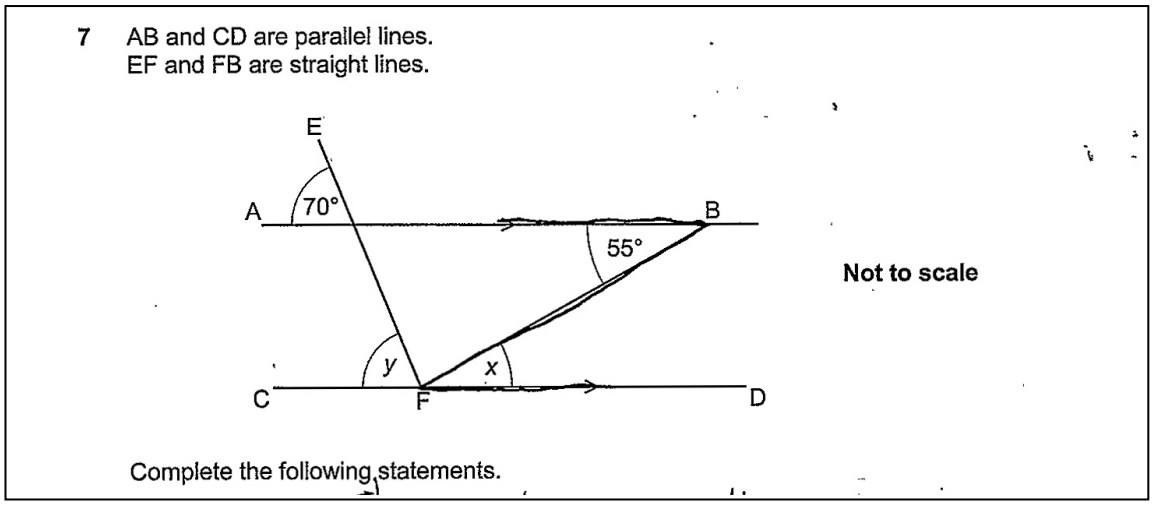

Figure 16: Example of candidate annotations suggesting use of "alternate angles" rule, M01 (Maths, Foundation tier)

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.

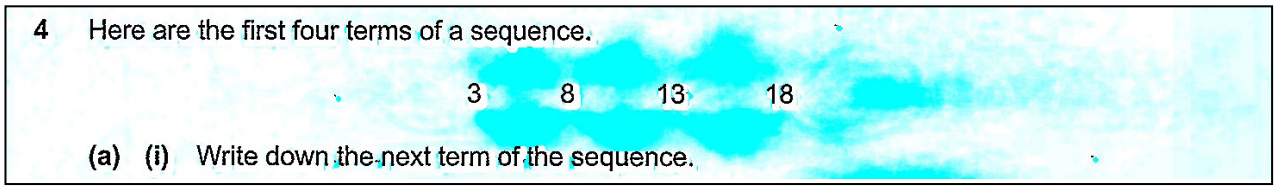

Figure 17: Heat map showing annotation on item M06 (Maths, Foundation tier)

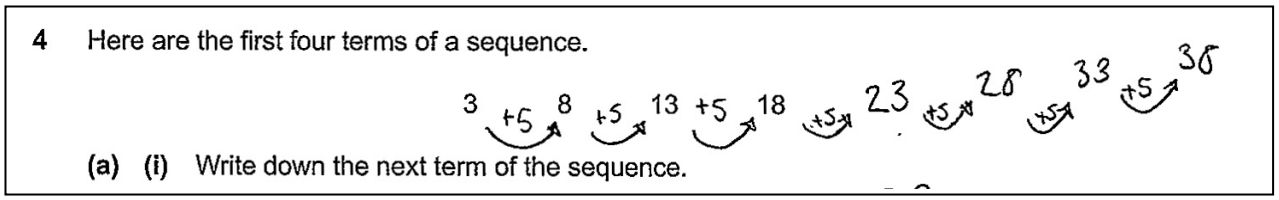

Figure 18: Example of candidate annotations showing addition between terms and sequence continuation, M06 (Maths, Foundation tier)

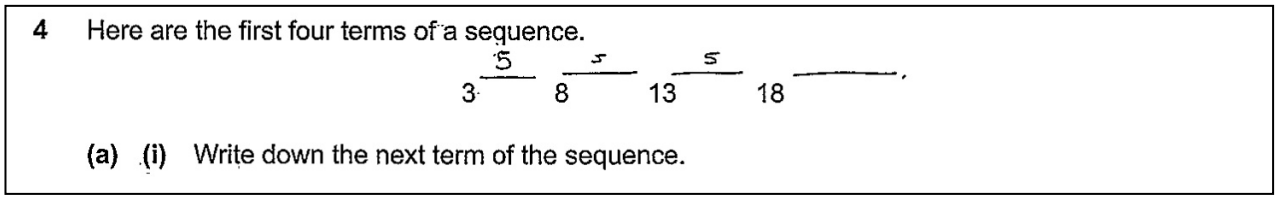

Figure 19: Example of candidate annotations marking on the differences between sequence terms, M06 (Maths, Foundation tier)

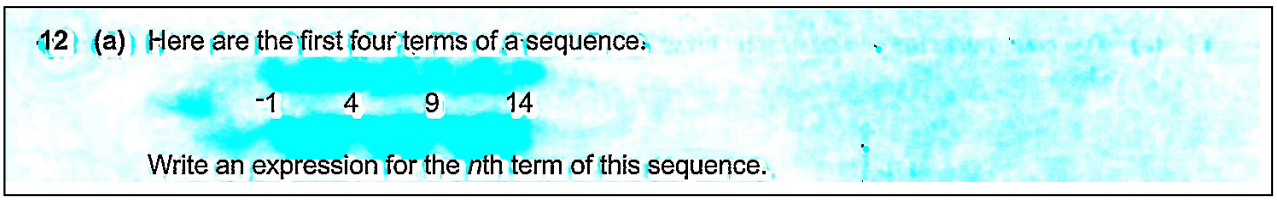

Figure 20: Heat map showing annotation on item M09 (Maths, Higher tier)

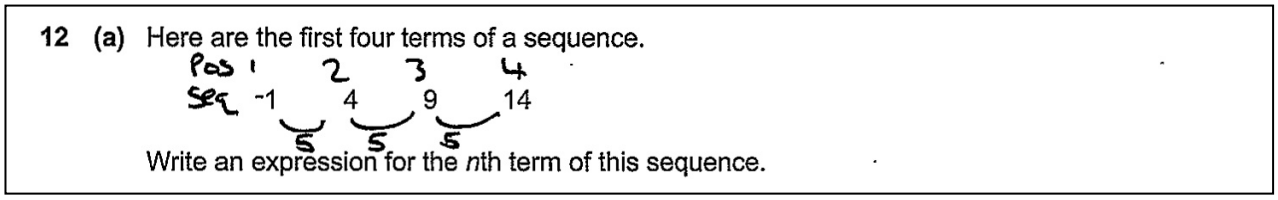

Figure 21: Example of candidate annotations numbering the sequence terms and marking on the differences between terms, M09 (Maths, Higher tier)

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.

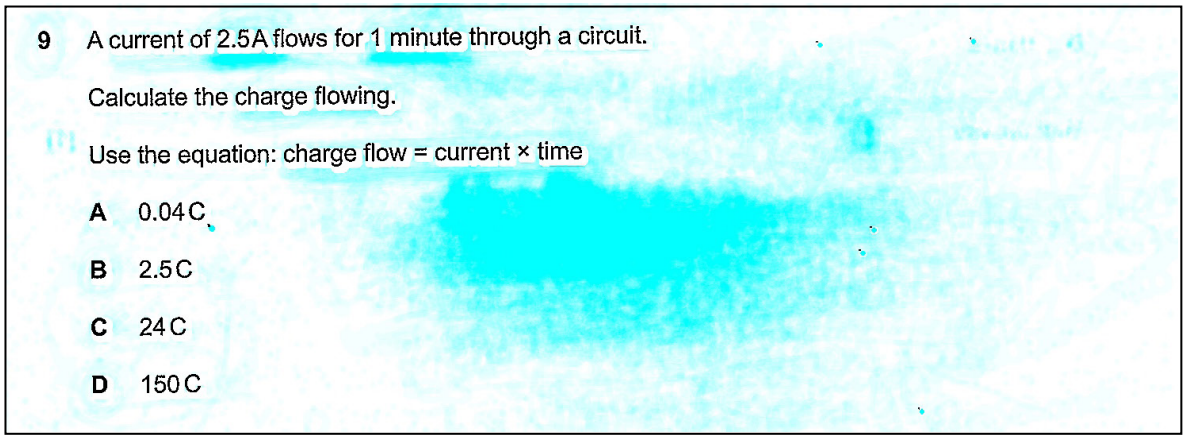

Figure 22: Heat map showing annotation on item P05 (Physics, Higher tier)

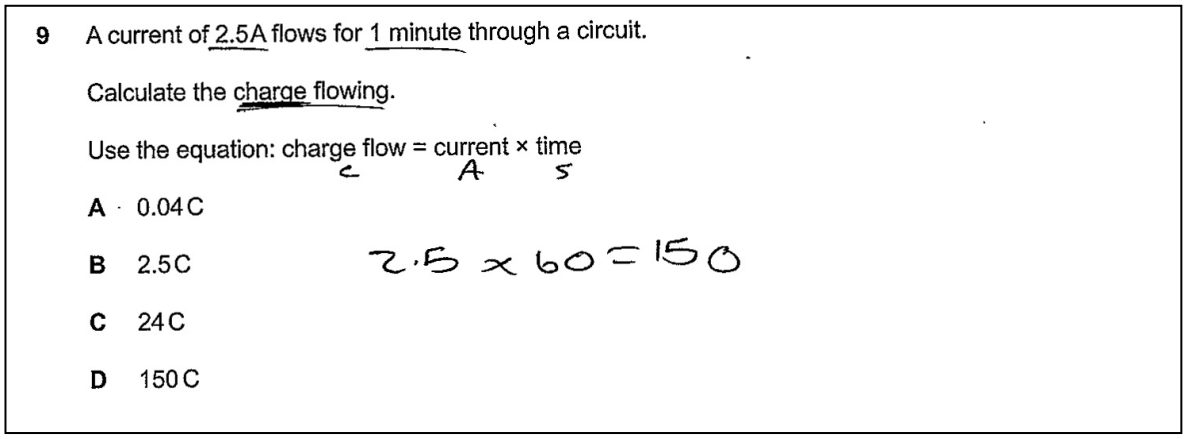

Figure 23: Example of candidate annotations including underlining of key information and calculation in white space, P05 (Physics, Higher tier)

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)

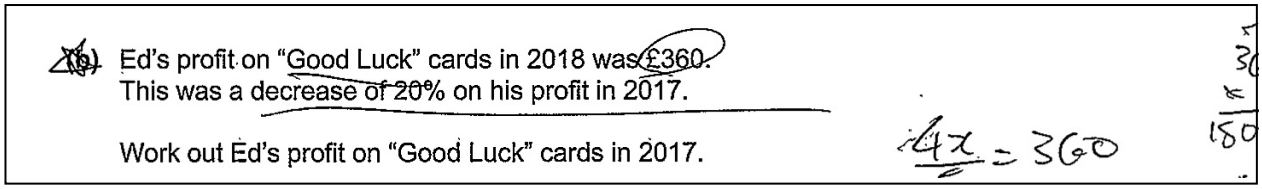

Figure 25: Example of candidate annotations including "overspill" from designated response space, M03-H (Maths, Higher tier)

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