Research and Analysis

A study of hard-to-mark responses

Why is there low mark agreement on some responses?

ofqual

Authors

This report was prepared by Caroline Morin, Beth Black, Emma Howard and Stephen D Holmes of the Strategy, Research and Risk Directorate

Acknowledgements

We would like to thank the exam boards for supplying scripts, and the examiners for conducting the marking and taking part in the meetings and discussions.

Contents

Executive summary

Black and Newton (2016) proposed a typology of marker disagreement and suggested 4 different categories of possible sources of disagreement – procedural error, attentional error, inferential uncertainty and definitional uncertainty. Procedural errors are errors that could be avoided if the correct procedure was followed (for example not marking all the pages of a paper). Attentional errors happen when examiners do not pay enough attention when marking (for example misreading part of a response). If examiners had paid sufficient attention, the correct mark would have been awarded. Inferential uncertainty arises when examiners have insufficient evidence to reach a definitive judgement (when interpreting the meaning of the candidate's response). Finally, definitional uncertainty occurs when the definition of the attainment construct (ie the mark scheme) and its scale is insufficiently precise to arbitrate between different views when determining the value of the candidate's response. The first 2 categories (procedural and attentional error) can be described as errors while the last 2 categories (inferential and definitional uncertainty) are present in responses which may have more than one legitimate mark.

This paper presents the results of a study aimed at testing these categories and identifying sub-categories or different instances of each type of error and uncertainty. This is the first attempt to validate the theoretical taxonomy suggested by Black and Newton (2016). In order to achieve this objective, we recruited 7 groups of experienced examiners (one group from each of 4 exam boards in English language and one group from each of 3 exam boards in biology). Each group comprised examiners on a single unit[1] who had all taken part in both the original marking and the post-results review of the marking process in summer 2016. Reviews of marking are carried out when a school believes that there was an error in the mark awarded to a candidate. When carrying out reviews of marking, examiners are asked to review the original marking of the script and to change a mark only when the mark scheme has been applied incorrectly. Scripts from the reviews of marking were chosen as they were more likely to include difficult-to-mark items. It should be kept in mind that these are likely to not be representative of the majority of scripts marked.

For each examiner group, we obtained 2 versions of a set of 100 scripts that had been through the post-results review process at the end of summer 2016: these versions were the original clean unmarked script, and the marked and annotated version that was reviewed. Each examiner marked 50 clean scripts and reviewed 50 annotated scripts. It was thought that having 'prime marks' as well as 'review marks' for these scripts might help us to tease out marks which represented a legitimate difference of marker judgement (potentially resulting from, for example, inferential or differential uncertainty), and marks which represented marking error (resulting from, for example, procedural or attentional error).

The marks awarded to each item within the scripts were analysed and items were selected where different patterns of mark agreement/disagreement arose. For example, where the modal mark awarded by examiners who marked was different from those who reviewed or when all examiners who marked or reviewed gave a

[1] Consisting of one exam paper

range of different marks. On this basis, we categorised the responses as 'hard-to-mark' responses.

During a series of one-day meetings, examiners who had been involved in the marking and reviewing were asked to look at a selection of responses that were hard-to-mark and to identify characteristics (such as response, item or mark scheme characteristics) and/or reasons for why a range of different marks had been awarded.

The results showed that the 4 typologies of marker disagreement appear to be exhaustive as all responses were successfully categorised using the 4 categories. In biology, most of the responses were classified as inferential uncertainty whilst in English language, the majority was classified as definitional uncertainty. A number of sub-categories were also identified, and there were some similarities and differences between the two subjects.

1 Introduction

In the English general qualifications system, there has been a desire to assess skills and knowledge using assessment methods that comprise questions requiring mostly constructed-responses, including extended pieces of writing. These are harder to mark than objective items, but are generally considered to be the most valid way to assess the higher-order skills that are valued. Marking those types of question consistently is therefore a very important aspect of the reliability of these assessments. Therefore, anything which can improve the reliability of the marking of these items would make a valuable contribution.

Marking can sometimes be difficult. This is evidenced by the fact that the same response can sometimes lead to a number of different marks awarded when marked by a number of examiners. We call these 'difficult to mark' responses and these will be central to the rest of this paper.

Black, Suto and Bramley (2011) proposed a framework (see Figure 1) outlining the different features that can impact on marking consistency. The framework includes 3 main types of features: item features, mark scheme features and the examinee response features. Item features include for example, the tariff of the question, item type, the size of the area for response, etc. The mark scheme features include whether the mark scheme is point-based or level-based and whether wrong answers are specified in the mark scheme. Finally, the examinee response features include handwriting, spelling, the typicality of the response, to name just a few.

Cognitive marking strategies are also an important part of the framework. Suto and Greatorex (2008) identified 5 different cognitive strategies used in marking: matching, scanning, evaluating, scrutinising and no response. These can also be mapped against the dual processing model (Kahneman and Frederick 2002; Stanovich and West 2002). The model distinguishes 2 different systems: system 1 which is automatic and unintentional and system 2 which is slow, deliberate and rules-based. Matching, scanning and 'no response' (confirming that no responses is present) would use system 1 while evaluating and scrutinising would use system 2. The cognitive strategies used to mark a response also interact with the 3 types of features described above. Together, these aspects will determine whether an item is hard or easy to mark.

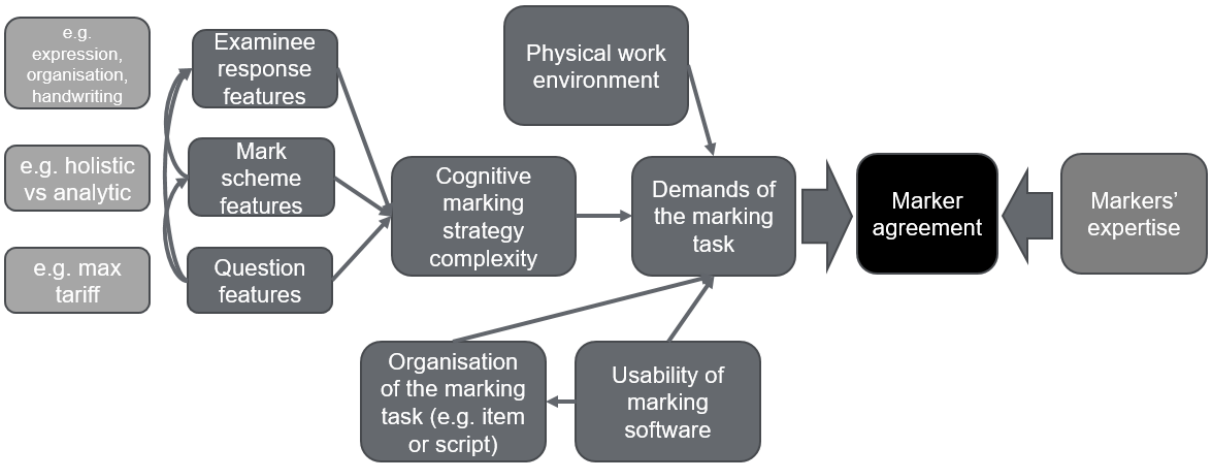


Figure 1. a version of Black, Suto and Bramley's framework (2011).

1.1 Typology of marker disagreement

Recently, Black and Newton (2016) proposed a typology of marker disagreement and suggested 4 different categories of possible sources of disagreement – procedural error, attentional error, inferential uncertainty and definitional uncertainty. This is, to our knowledge, the only attempt at categorising the possible sources of marker disagreement.

Procedural errors are errors that could be avoided if the correct procedure was followed (for example, do not mark all the pages of a paper). Attentional errors happen when examiners do not pay enough attention when marking, resulting in misreading part of a response for example. If an examiner had paid sufficient attention, the correct mark could have been awarded. Inferential uncertainty arises when examiners have insufficient evidence to reach a definitive judgement (for example, in interpreting the meaning of a candidate response). Finally, definitional uncertainty occurs when the definition of the attainment construct and its scale is insufficiently precise to arbitrate between different views.

Table 1. Different types of errors and uncertainties and a short definition for each

Procedural error	Markers make mistakes and do not follow procedure eg do not mark all pages of a response or apply the wrong mark scheme	error
Attentional error	Markers have concentration lapses eg mis-key a mark or misread a critical word/number	error
Inferential uncertainty	Markers have insufficient evidence to reach a definitive judgement. Different markers award different marks on the basis of different inferences.	uncertainty
Definitional uncertainty	Markers' views differ on the definition of the construct and its quality scale ie quality means (subtly) different things to different examiners. The mark scheme is insufficiently precise to arbitrate between different views.	uncertainty

The main objective of this study was to determine whether the 4 categories of errors and uncertainties suggested by Black and Newton (2016) were exhaustive. A further objective was to identify sub-categories or different instances of each type of error or uncertainty, if present.

The study involved groups of examiners marking and reviewing the same scripts from 3 biology units and 4 English language units. Once the scripts were marked, individual responses were chosen for discussion, based on the spread of marks awarded in the marking and reviewing exercises. These responses were discussed in a one-day meeting and the discussions were used by the researchers to classify each response.

2 Method

2.1 Materials and participants

The 2 subjects and the individual units were chosen because they had a large number of reviews of marking in summer 2016. This methodology was chosen given that scripts sent for a review of marking usually have more chance of containing one or more items that are difficult to mark. From all the scripts that were sent for a review of marking in summer 2016 for each unit, 130 were selected, with similar mean, median and range to all the scripts that went through a review of marking for that unit. These scripts were requested from the exam boards in 2 versions: cleaned scripts for marking and annotated scripts for review. Once received, they were anonymised and from the 130 scripts, 100 scripts were selected to be used in the study. Scripts that either used a scribe, had additional pages at the end or were not awarded any marks were excluded. From the remaining scripts, 100 were selected randomly.

We also requested the scripts that were used during standardisation so that the live standardisation process[2] could be replicated in the study and therefore remind examiners of the marking standard and mark scheme before embarking upon the experimental marking and reviewing.

We provided the exam boards with recruitment emails to send to suitable examiners. For each unit, a principal examiner (PE)[3], and 7 team leaders who had taken part in the reviews of marking the previous summer were recruited, where possible. Examiners interested in taking part in the study contacted us directly or contacted the exam boards, as specified by each exam board. When exam boards' preference was for examiners to contact Ofqual, the selection was made on a "first come first served" basis until all positions were filled within a unit.

[2] The standardisation process is either a face-to-face meeting or online exercise where examiners learn to apply the mark scheme through practicing marking candidate responses.
[3] The Principal Examiner (PE) is the most senior examiner and is the one who sets the standard for the unit. Team leaders are senior examiners who oversee the marking of a group of around 6 to 8 examiners.

2.2 Procedure

2.2.1 Standardisation

Both the standardisation and main marking were carried out online using a bespoke marking system. Once examiners had completed the standardisation scripts for the unit to which they were allocated, the marks awarded by the examiners were sent to the PE. The PE then compared the marks awarded by the examiners to the definitive mark for each question (ie the mark, determined in advance by the PE). They were then asked to have discussions with the examiners to ensure that their marking met the standard required during live marking. Once this was done, the examiners were allowed to start marking the 100 scripts.

2.2.2 Marking and reviewing

The marking exercise was split in 2 parts: 50 original clean unmarked scripts to mark and 50 marked and annotated scripts to review (with original marks and annotations). The 100 scripts were randomly divided into 2 batches of 50 scripts; batch A and batch B. Half of the examiners marked batch A and reviewed batch B whilst the other half marked batch B and reviewed batch A. In each set of 50 scripts, scripts were presented in a random order to avoid sequence effects. When reviewing the scripts, examiners had to decide whether the original mark awarded represented a legitimate mark (no change) or an error (change the mark). Reviewing these scripts allowed the identification of responses where there were genuine errors in the original marking, rather than some form of legitimate difference of professional opinion.

2.2.3 Software

Unlike the marking system used by the exam boards, the bespoke system did not allow for the monitoring of examiners' performance using seeds. The main advantage of the bespoke system was that no examiner would be advantaged/disadvantaged as it was a new system to all examiners as the exam boards use different systems. Whole scripts were loaded onto the system and examiners could mark either by question or by script. One important difference between the bespoke system used in this study and the systems the exam boards use is that there was no access to script annotation tools.

2.2.4 Analysis – response selection

Once all the marking and reviewing was completed, the marks awarded were retrieved from the marking system and analyses looking at the mode, mean and spread of the marks awarded were carried out. Different patterns were selected for subsequent discussion, for example, where the modal mark for marking and reviewing was different or where all examiners had given a different mark and review mark. The original marks awarded at live marking as well as the live review marks were also available. The chosen responses were printed on paper and presented for discussion during the one-day meetings with the examiners. A total of 64 responses were discussed during the 3 biology meetings and 56 during the 4 English language meetings.

2.2.5 Examiners' meetings

Each meeting started with a presentation describing the 4 categories of disagreement with some examples. Examiners were then asked to look at a number of responses one at a time and discuss why the marking may have led to disagreement. The objective of the meetings was to collect examiners' view on how the marks awarded in the study can be explained. The meetings were audio recorded and later transcribed to help the coding of each response by the researchers.

2.2.6 Coding

Three researchers had a one-day standardisation meeting for each subject in order to agree on how to code the responses. During that meeting, a number of example responses were discussed, with reference to the audio recordings of the meetings and their transcriptions. By the end of the meeting, all researchers felt confident that they knew how to code the responses into the 4 categories of error and uncertainty or in new categories if none were applicable.

Two of the researchers then both coded all the responses based on the discussion that took place during the one-day meetings using the audio recordings, the responses and the transcription of the discussions during the meetings. They then met and compared the category/categories attributed to each response. When the 2 researchers did not agree, the responses were discussed with the third researcher who would adjudicate. In depth discussions between the 3 researchers also took place in order to establish the sub-categories presented in the following sections and on occasion this meant revisiting previous codings to ensure consistency.

3 Results

The results for biology and English language are presented separately as there were substantial differences between the coding of the responses for each subject. Within each subject, the 4 high level categories are considered in turn and any additional subcategories developed/discussed. When examples are used, the candidate's response and the mark scheme will be presented as well as an explanation of how the response exemplifies that sub-category.

3.1 Biology

3.1.1 Number of responses

In biology, a total of 64 responses were discussed across the 3 units. Each response could be classified as having characteristics of one or more of the 4 categories of error and uncertainty. The Venn diagram below presents how the 64 responses were categorised.

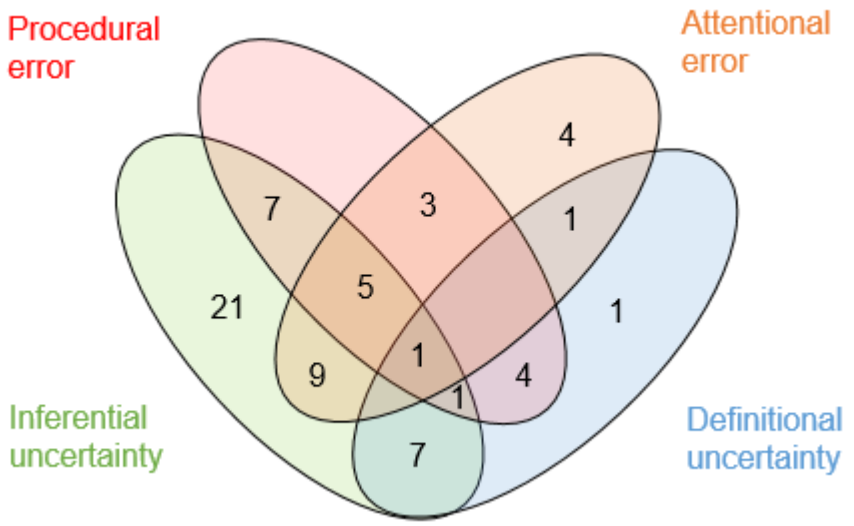


Figure 2. Categorisation of the responses discussed in the meetings for the 3 biology units. NB: responses categorised in 2 or more categories can be instances of the different categories interacting or different categories affecting different parts of the response independently.

The first objective was to see whether Black and Newton's (2016) categories were exhaustive or whether some of the responses would not fit in any of the 4 categories. The 64 biology responses could be categorised using the framework in one or more categories (see Figure 2). Most of the responses in biology appeared to have caused, at least in part, inferential uncertainty.

The second objective of the study was to try and describe sub-categories or different instances exemplifying each category. This will be covered in the next sections for the biology responses.

3.1.2 Procedural errors

Procedural errors are those that arise when an examiner does not follow the procedure. In biology, 21 of the 64 responses were classified in this category and all of them fell into one sub-category.

Clear misapplication of the mark scheme

In all instances, the PE and others clearly stated in the meeting that some marks were outside the range of acceptable marks even taking into account legitimate difference of opinions. In other words, an unacceptable application of the mark scheme or a lack of application of the mark scheme. In the study, when markers awarded a mark that fell outside the acceptable range for a response, as defined by the PE and others, this was classified as a procedural error as it would unambiguously contravene the mark scheme.

> *In this response, one examiner marked the response as a 2 and one examiner reviewed the response as a 3.*
>
> - So it could be one or zero.

- I can see why someone's given it 2 about the poachers but they shouldn't have done.

3.1.1 Attentional errors

Attentional errors are those that arise because of lapses of attention by examiners. There is no ambiguity in the response but an attentional slip means they have not processed and interpreted the response correctly. It was the second most frequent category for biology, with 23 out of 64 responses classified as having appeared to cause some attentional errors.

Inattention resulting from word spotting

The majority of responses in this category in biology are attentional errors where the examiners have done some form of word spotting. This is in line with the use of matching and scanning as a cognitive strategy as described earlier in this paper. Examiners scan the responses trying to match keywords present in the mark scheme with words present in the response. It sometimes happens, as in the example in Figure 3, that scanning for words, and ignoring the context of that word in the sentence, results in a mark being rewarded erroneously. In this example, the mark for 'vasoconstricts' (highlighted in yellow) should not have been awarded in the context of the response, as the candidate has said that the blood vasoconstricts rather than the blood vessels vasoconstrict. Scanning or 'word spotting' as a marking strategy has led to inattention to the rest of the responses, resulting in a marking error – in this case the over-rewarding of a mark. Word spotting could also lead to under-rewarding of a mark where examiners fail to spot material that is creditworthy because their attention is captured elsewhere in the response.

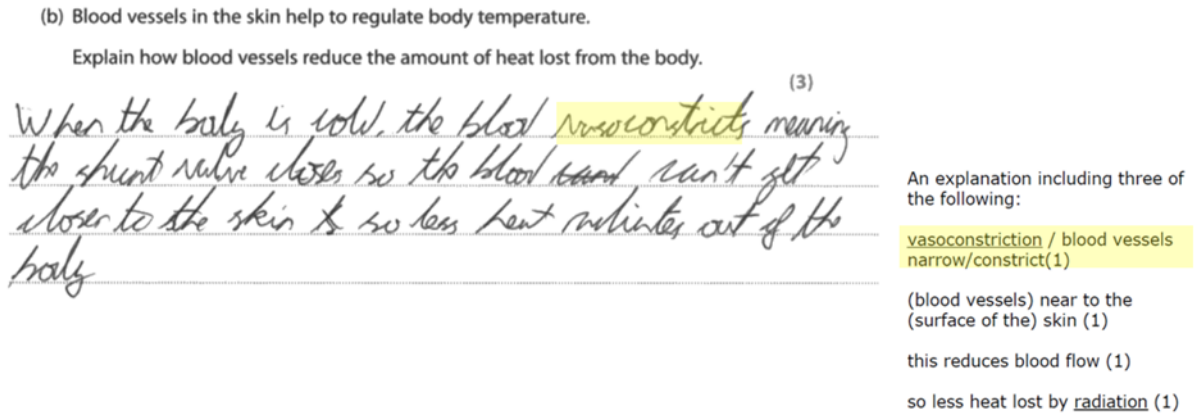


Figure 3. Example of inattention resulting from word spotting and leading to marking error.

3.1.2 Inferential uncertainty

Inferential uncertainty is when examiners have insufficient evidence in the response to reach a definitive judgement. The key issue is in deciphering what the candidate knows or intended to convey when they produced an ambiguous response. In biology this often revolves around using terminology that is not quite aligned with the terminology in the mark scheme, but as the examples below show, there are a variety of sources of uncertainty. Most of the responses (51 out of 64) were

categorised as having appeared to cause some inferential uncertainty. This category describes some situations where there are legitimate differences of opinion, whereby neither opinion seems unreasonable given the lack of unambiguous information with which to arbitrate between those opinions.

Differing interpretations of handwriting – identification of words/letters

At the most basic level of meaning extraction, inferential uncertainty involves deciphering what the candidate has produced in a response in terms of letters or words. This involves making an inference in order to be able to award a mark. In the example shown in Figure 4, identification of whether the letter is upper or lower case (F or f) has led to different inferences being made by examiners – and this is pivotal for the nature of this question and whether the candidate understand the difference between dominant alleles (symbolised using upper case letters) or recessive alleles (lower case). The marks awarded in the study ranged from 0 to 2. This means that not all examiners extracted the same information (F or f) and hence examiners gave the response a range of marks based on the information they extracted.

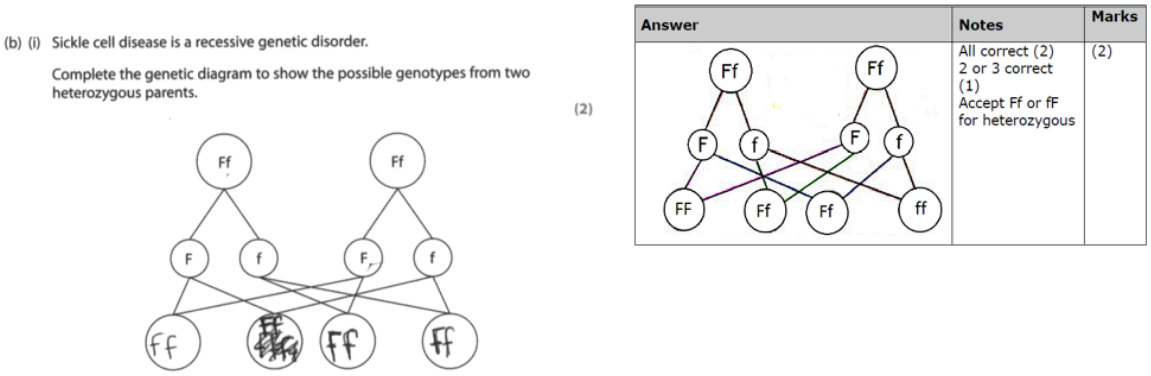
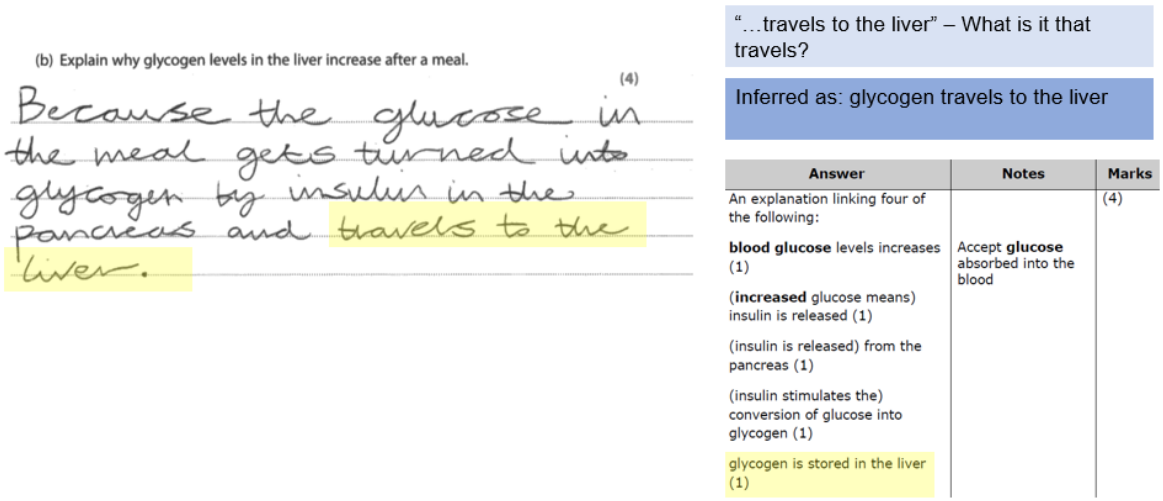


Figure 4. Example of a response involving the extraction of information at the lowest level (letters).

Differing interpretations of sentence/paragraph meaning

Whereas the previous category was centred around identification of single letters, this category is looking at interpreting sentences and paragraphs. In the framework presented earlier, this is linked to the system 2 processes, more specifically evaluating and scrutinising a response. In the example shown in Figure 5, the extraction of meaning at word-level was fairly straightforward but the fact that the candidate wrote the answer as one sentence, without punctuation, made extracting meaning very difficult. In this case, the candidate talks about a substance that travels to the liver (highlighted in yellow) but it is unclear which substance it is, given the length of the sentence and the lack of punctuation. The correct answer is glycogen but examiners needed to make a decision as to whether the response indicated that it was the glycogen, the glucose or the insulin that travels to the liver. Given a range of marks were awarded, we can conclude that not all examiners inferred the same response. It is also worth noting that the phrase 'travels to the liver' is also not quite the mark point, given that the mark point is about storage. So there is some additional ambiguity about whether 'travel to' is actually 'storage'. Overall, different examiners have made different inferences and determinations

about the extent to which the candidate understands this process; certainly, there is ambiguity in the way the response, and more specifically the sentence, has been worded.



(b) Explain why glycogen levels in the liver increase after a meal.

(4)

Because the glucose in the meal gets turned into glycogen by insulin in the pancreas and travels to the liver.

"…travels to the liver" – What is it that travels?

Inferred as: glycogen travels to the liver

Answer	Notes	Marks
An explanation linking four of the following: **blood glucose** levels increases (1) (**increased** glucose means) insulin is released (1) (insulin is released) from the pancreas (1) (insulin stimulates the) conversion of glucose into glycogen (1) glycogen is stored in the liver (1)	Accept **glucose** absorbed into the blood	(4)

Figure 5. Example of a responses involving meaning extraction at a higher level than letter or word-level.

Differences in making inferential leaps (minor)

When extracting meaning from a response, there may be a small gap in literal meaning of the response as it is actually given and the meaning conveyed in the mark scheme. On these occasions, some examiners might make small inferences that the response meaning is close enough to the target meaning in the mark scheme; that, effectively, the candidate knows the material; whereas others may not make this inference. Figure 6 gives an example of this. In this example, the word "mindset" was accepted by some as a substitute for "psychological effects", but not by others. This was evidenced by the large range of marks awarded for that response. These small inferences are examples of "benefit of the doubt"[4]. This concept of "benefit of the doubt" could be seen as an explicit acknowledgement on behalf of an examiner that the response wording/meaning is not quite a good enough match to the wording/meaning in the mark scheme.

> *Examiner a*
>
> I think it's the word mindset. I think that people have decided whether mindset is worth…
>
> *Examiner b*
>
> Oh right. That could be psychological
>
> *Examiner a*

[4] 'Benefit of the doubt' (or 'benefit of doubt'/'BOD') - while this may be sometimes legitimately applied in live marking, this should not be the case in reviewing marking (in ROMMs). 'Benefit of doubt' should not be used to decide that an original mark was incorrect because this would be effectively replacing one legitimate mark with another legitimate mark.

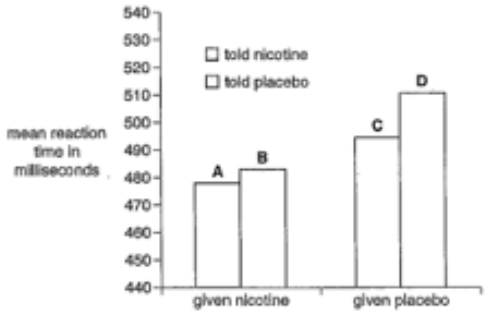
> So that could have got 2 or 4 depending on whether they've decided on mindset.

Look at the graph.

It shows the results from a trial where four groups of people were tested.

- Group **A** was given nicotine and told it was nicotine
- Group **B** was given nicotine and told it was a placebo
- Group **C** was given a placebo and told it was nicotine
- Group **D** was given a placebo and told it was a placebo

Each group's mean (average) reaction time was then recorded.



Suggest why the trial was designed in this way, and explain what the results show.

The quality of written communication will be assessed in your answer to this question.

The trial was designed this way to find more information about how nicotine works with the participants mindset. The results show that giving a group nicotine, ~~whether~~ either ~~tellin~~ telling them it was nicotine or a placebo will decrease reaction time more significantly compared ~~to both told nicotine and told placebo~~ groups C and D. Giving nicotine is more effective than giving a placebo. [6]

Design

- to test the effects of a placebo on reaction time
- placebo effect when given/not given any nicotine
- need to use a method to make sure that human mind / psychological effects / bias has not influenced results
- placebo used to compare effectiveness of nicotine / show the effect of nicotine

Results

- faster reaction times with nicotine / ORA
- faster reaction time if they think they have nicotine
- placebo has an effect on results
- placebo has less of an effect than nicotine
- results show that nicotine has a bigger effect than placebo even when people are told they've been given a placebo

Figure 6. Example of a small inference.

Differences in making inferential leaps (major)

Examiners also suggested that marks for some responses were evidence of large inferences whereby the gap between the meaning in the response, and the meaning in the mark scheme was larger. Sometimes, in these circumstances, the participants in the study referred to examiners as 'doing a lot of work on behalf of the candidate' by assuming a lot of knowledge and understanding behind their ambiguous answer. Figure 7 is an example of a large inference.

This question is about why the glycogen level in the liver of the male with untreated diabetes would be different from the healthy male after a meal. There seems to be confusion because the candidate switches from talking about the man with type 2 diabetes to talking about the healthy man. The last sentence states:

"Glycogen levels would be a lot lower for the healthy man" (highlighted in yellow in Figure 7).

The inference in this case is that examiners assumed that the candidate meant that the glycogen level was lower than for the healthy man. This inference is large because by inserting the word 'than' essentially creates the opposite meaning of the sentence. In other words, while there is no unambiguous evidence that the candidate knows the correct answer, some examiners have inferred that they do, possibly from the rest of their answer which is phrased in terms of the man with diabetes.

> He had 'the glycogen level would be a lot lower than for the healthy man'. You're only missing 'than', but it changes the whole context.

This response was a particularly problematic one with examiners awarding the whole range of marks. Like the example in Figure 5, this answer also suffers from very long sentences and limited punctuation which may lead to different inferences being made. Also, words have been inserted with an arrow and some words are difficult to decipher. All these characteristics have had a different impact on each examiner's attempt to extract meaning and how they have awarded different marks accordingly.

(c) A male with untreated Type 2 diabetes ate lunch with the same carbohydrate content as the healthy male in the graph.

Explain why the glycogen level in the liver of the male with untreated Type 2 diabetes would be different from the healthy male after this meal.

(3)

If you have type 2 diabetes then your body and pancreas may become resistant to insulin as well as cells, normally this would be treated by insulin injection however because he is untreated it may be that insulin cannot be secreted by the pancreas meaning too much glucose in his body which then is stored as glycogen. So the glycogen level would be a lot lower for the healthy man.

Question number	Answer	Notes	Marks
2 (c)	An explanation linking **three** of the following points: glycogen levels lower / graph would be flatter (1) a person with type 2 diabetes does release insulin / the amount of insulin released is not enough (1) but cells have become resistant to insulin (1) so no / less glucose is converted to glycogen (1)	Accept cells do not respond to insulin	(3)

Figure 7. Example of a large inference.

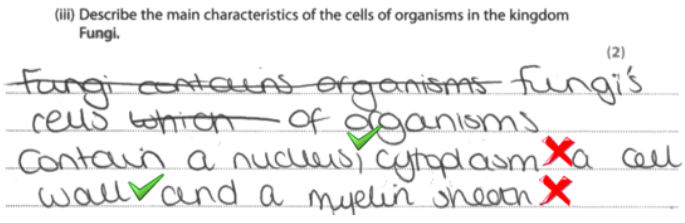
3.1.3 Definitional uncertainty

Definitional uncertainty is where there is a lack of precision in the definition of the construct and which therefore means that examiners' views can differ on how to interpret the quality scale. This category describes a legitimate difference of opinion, whereby neither opinion seems unreasonable given the unavoidable imprecision of the quality scale in the mark scheme. This is (usually), at least to some extent, unavoidable because it is not possible, particularly in levels of response mark schemes, to explicitly arbitrate for every single possible permutation of quality of a number of skills exhibited in a response. Fifteen of the 64 responses were classified as having been partly or wholly caused by definitional uncertainty.

Mixed responses/differing decisions based on the mark scheme

One sub-type of definitional uncertainty was related to mixed responses. In biology, mixed response examples included those where there was some correct and some incorrect information. There are further subordinate categories: in some instances, the incorrect information is unrelated to, or has no direct bearing on the material which is correct (see Figure 8); whereas in some cases the incorrect information is contradictory to other (otherwise correct) parts of the response (see Figure 9). Some mark schemes are explicit about what to do in some instances[5] – thus avoiding this source of definitional uncertainty (though still susceptible to procedural error if the examiner ignores it).

Figure 8 shows an example of a mixed response containing incorrect information that has no bearing on the previous correct information. In this case, the candidate has correctly identified 2 characteristics of cells in fungi but also added 2 incorrect but not contradictory characteristics. Some examiners have ignored the unrelated material and some have negated a mark awarding either 1 or 2 marks during the study. The decision concerning what to do with unrelated incorrect information seems to be different from examiner to examiner. For the examiners who negated marks, this may be because they thought that the candidate was essentially guessing, listing any cell feature they could think of, whether or not relevant to fungi, and the mark scheme did not legislate what to do is such a situation.



Question number	Answer	Notes
6 (a) (iii)	A description including the following points: cells have cell walls (1) cells do not have chlorophyll / chloroplasts (1) (organised into) mycelium / hyphae (1) cells have nuclei (1)	Accept: cell wall not made of cellulose / is made of chitin (2)

Figure 8. Example of a mixed response in biology containing correct and unrelated incorrect information.

There were also instances where mixed responses would include correct information and incorrect information contradicting correct information elsewhere. In the example presented in Figure 9, the information highlighted in red suggests that the process described in the first sentence may not be fully understood as the nitrogen fixing bacteria is part of the fixation cycle (see mark scheme). Not all examiners agreed as some have awarded full marks to the response. Again, examiners differ in how they handle contradictory incorrect information.

> *Examiner a*

[5] For example, one of the mark schemes specifies that if a question asks for 2 items, examiners should stop marking after the second item mentioned.

> If you do lots of ignoring of wrong things there might be 6 marks
>
> *Examiner b*
>
> If you start penalising everything that's wrong you can just about fall into 2 [marks]

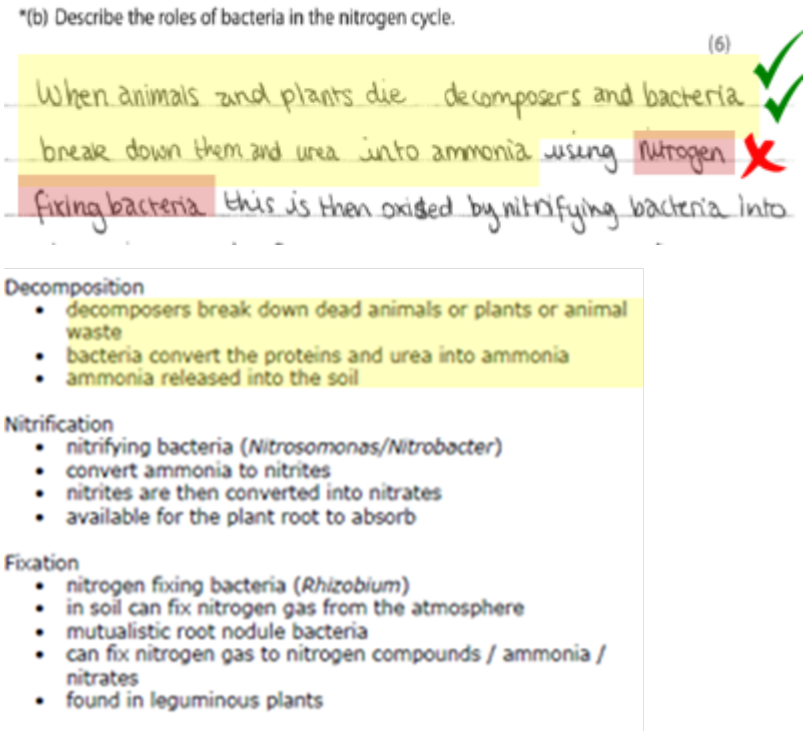


Figure 9. Example of a mixed response in biology containing correct and contradictory incorrect information.

3.2 English language

3.2.1 Number of responses

In English language, a total of 56 responses were discussed across the 4 units. Each response could be classified as having characteristics of one or more of the 4 categories of error and uncertainty. The Venn diagram below presents how the 56 responses were categorised.

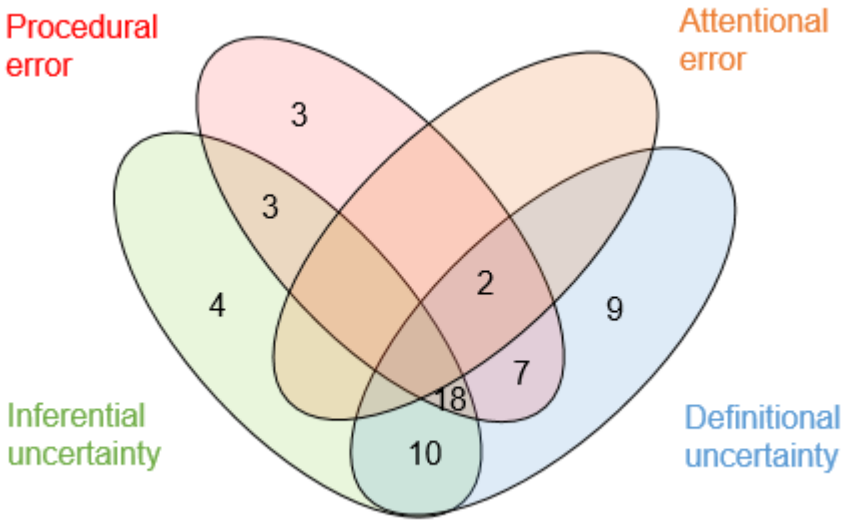


Figure 10. Categorisation of the responses discussed in the meetings for the 4 English language units. NB: responses categorised in 2 or more categories can be instances of the different categories interacting or different categories affecting different parts of the response independently.

The first objective was to see whether Black and Newton's (2016) categories were exhaustive and as can be seen in Figure 10, all the English language responses could be categorised using the framework in one or more categories. Most of the responses in English language contained, at least in part, definitional uncertainty (48 of the 56 responses).

The second objective of the study was to try and identify sub-categories and describe different instances exemplifying each category. This will be covered in the next sections for the English language responses.

3.2.2 Procedural errors

Procedural errors are errors that could be avoided if the examiner followed the correct procedure. In English language, we encountered 15 instances of procedural errors falling into 3 sub-categories.

Clear misapplication of the mark scheme

As described in the biology results section, this sub-category of disagreement arises when an examiner awarded a mark that was clearly stated in the meeting as being outside the range of acceptable marks (even taking into account legitimate difference of opinions). These instances unambiguously contravene the mark scheme and are classed as procedural errors based on the fact that if the examiner had followed the information on the mark scheme, he/she would have awarded a mark within the acceptable range.

Applying the wrong generic mark scheme

The second sub-category is linked to the use of 2 generic mark schemes (both with the same number of levels) but for 2 different questions and with different maximum marks. For each of the 2 questions, the range of marks available in each band of the

mark scheme would be different. If an examiner did not use the mark scheme for the correct question, he/she may award marks in the correct band (for example 'level 2') but given the range of marks would be different, the mark awarded would be incorrect. This was supported by the annotations of the original marker, which indicated that the marker had identified the correct band, but then selected the mark from the wrong mark scheme, resulting in the wrong mark.

Ignoring instructions around the independence of 2 marks

The third sub-category is where examiners ignore rules around how different marks for the same response should be awarded independently of one another. This could happen, for example, where an essay is marked for content, and is also marked for the quality of written communication; or where 2 different assessment objectives' marks are awarded for the same response. Some markers should make the 2 appraisals such that, depending on the response, it is possible that one is high on its scale and the other is low. However, some examiners make a false assumption that both marks need to be proportionate or be in the same band.

In the example given below, senior examiners were definitive that this is something that assistant examiners are told during the standardisation process which is why this sub-category was classified as a procedural error. Here is an extract of the discussion surrounding one of the responses:

> - I think this one raises an issue, you see, and I think it's because you're thinking well the [sentence structure, punctuation and spelling] on this is modest. I mean there are a lot of errors again.
> - And how far on the content can I go above the band I've allocated it to? Because when you start pushing up to 10s and 11s there's a big discrepancy now appearing. And you may have been warned that this doesn't happen very often. It can happen, it doesn't happen very often. And I think then maybe you've got to have a bit of courage maybe to go with that difference. So I just wonder if that is an issue, would be an issue on this one for examiners. How far you go on the content, and it is a big question.

For some boards the mark schemes for the 2 assessment objectives are presented side by side on one page which may suggest that they should be awarded marks from the same band although examiners are told it is not necessarily the case during the standardisation meeting.

> - You want it to be as easy as possible to see it [mark scheme for the 2 assessment objectives], and that's I think why it's always been traditionally on the same page, to make it easy to look. There is that thing that if it's on the same page, if it's parallel, that's what we're expected to do, and I'm sure a lot of markers think that.
> - Yeah, but we're told at standardisation weren't we?
> - I know.

> - Until we're blue in the face.

3.2.3 Attentional errors

There were 2 instances of attentional errors in English language.

Inattention leading to missed information

This sub-category is different from the one identified in biology (inattention resulting from word spotting) where the main cognitive processes involved were matching and scanning. In the current sub-category, the main cognitive strategy involved is "evaluating" the evidence, in other words, extracting the meaning. Some examiners did not pay enough attention when carrying out the evaluation process which meant they did not extract all the meaning and this led to an erroneous mark being awarded.

> I think because the points are quite densely packed, people will miss them as well, that's another issue with a script that's so tight like that.

This suggests that attentional slips are more likely to happen in scripts which are more difficult to parse; and that the way in which candidates present or format the same information may make it more or less susceptible to such errors.

This sub-category is different from inferential uncertainty – which is where there are differing and legitimate interpretations of what has been written. This category describes a situation where, if the examiner had properly read the answer, they would be able to award the correct mark.

3.2.4 Inferential uncertainty

As a reminder, inferential uncertainty arises when the response provides insufficient evidence to reach a definitive judgement, (for example, in interpreting the meaning of what a candidate writes) such that different examiners are likely to award different marks. In the case of English language responses, 35 of the 56 responses were classified as causing some inferential uncertainty. In this section, we will describe 4 sub-categories of inferential uncertainty that were observed in English language responses.

Heuristics based on characterisations of the candidate

Faced with relatively challenging inferences to be made, some markers may sometimes use rule of thumb inferences, or 'heuristics'. Heuristics are mental shortcuts that allows people to solve problems and make judgments quickly and often efficiently. In some of the instances in this research, it is possible that the heuristics have led to marks which are not fair representations of the quality of the response.

The first sub-type relates to inferences regarding the candidate based on the response that examiners are marking. It colours the evaluation of the response for some examiners and can affect marker agreement. It is important to note that such 'biases' are not systematic in that they are unlikely to affect all examiners, or affect

all examiners in the same way. Hence the responses received a range of marks. [6] Examples of heuristics observed included quick characterisations of candidates. Some examiners commented on their perception that, for example, some candidates appeared to be second language speakers of English, or that they were a foundation or higher student, or male or female candidate (depending on handwriting), and then applying a heuristic to arrive at a final mark.

> So that's probably something else that's perhaps unique to this one is that they're probably a more of a foundation style candidate, aren't they?

> - There's a lot going on there. Doesn't like to stretch himself too much, this is definitely a boy.
> - You're right it is a boy.
> - Scruffy writing, can't be bothered writing too much

<u>Heuristics based on superficial features</u>

Another sub-type of inferential uncertainty concerns inferences based on superficial features of the response. This is the case when examiners base their judgement to award marks on superficial features or when these features influence the marks they award. Here are some examples of what examiners said:

> I see football, I see band 4, straightaway.

> If you read none of these words, and you see they've done two-and-a-half sides, it's written in paragraphs, the handwriting's readable. And a candidate who gets 5 out of 8 for question one looks like that often.

<u>Subtlety</u>

Another sub-category is related to the subtlety of the piece. Classified as inferential uncertainty as potentially useful information on the response is not immediately available, and may need to be inferred. Some examiners may not see the subtlety on first reading while others do, which may lead to different marks being awarded. Here are a few examples of discussions around subtlety:

> I think this is another one though that the more you read it, the more you see in it.

[6] Where such techniques for making inferences are used by some examiners, this will likely result in a range of different marks being awarded, some of which might be classified as error or a misapplication of the mark scheme.

> I think it goes back, one of the points I added … subtlety and, you know, I think this isn't shouting from the tree tops this is an example of this, this is an example of that. But I think it does absolutely have that understanding and I think when we've often talked about a band 3 being very pedestrian and I don't think this is.

> - So the point with this script then from what you've said is that it appears at first glance to be rather pedestrian but actually on careful reading there is nuance in there. There is…
> - Yeah, subtlety.

Primacy/recency effect

This sub-category is perhaps a particular variation on mixed responses (see later) and relates to the fact that some examiners will be more influenced by either the beginning or the end of a response and their associated qualities. If the start of the response is bad and the end is good, some examiners will weight the beginning of the response more heavily than the end and give a more negative mark. Those examiners who are more influenced by the end of the response would potentially award more marks to the response. This may have been more prevalent in the current study as the examiners did not have access to the annotation facility they usually use when marking for the exam boards. This meant that they had to rely on their memory of the response a lot more, perhaps leading to different overall memories and hence different assessments (Aldrovandi, Poirier, Kusev and Ayton, 2015).

> - Especially if you've got a script where the beginning and the end are rather different, and your last impression is the end
> - Yeah. I'm wondering if the variation may have been caused by, I mean we know we can't remember everything, so unless you make notes, you may remember the positive more than the negative.

3.2.5 Definitional uncertainty

Most responses in English language were mixed responses with strengths and weaknesses in addressing different aspects of the marking points mirroring the correct and incorrect (irrelevant/contradictory) sub-categories in biology. When this is the case, examiners sometimes find it difficult to agree on a mark. It is interesting to note that for English language responses, inferential uncertainty also appeared to be present for more than half of the responses that were classified as having definitional uncertainty (in 28 responses out of 46). Sometimes the 2 types of uncertainties were independent, for example for different parts of the response. However, in a large number of them, inferential and definitional uncertainty were interacting. This meant that examiners would extract slightly different information

from a response (inferential uncertainty) and also had a slightly different way of understanding the construct that needed to be evidenced (definitional uncertainty). Depending on the alignment of the information extracted and the understanding of the construct to be demonstrated, different marks could legitimately be awarded by the different examiners.

Mixed responses/differing decisions based on the mark scheme

Many responses involved weighing up different qualities in the response, and while these various qualities expressed in the mark scheme, the overall articulation of the quality scale does not describe the particular profile of mixed/imbalanced qualities of the response. Sometimes this was just because of the descriptions in the band descriptors. Sometimes it was also because of a contradiction between 2 different parts of the mark scheme, for example, where the top band descriptors require the answer to be 'consistently focused on the question', while the general marking guidance suggests the use of "positive marking", ignoring incorrect or irrelevant material. The former could lead to mark disagreement as examiners may not agree on the "best fit" in terms of the band in which the response should be categorised. The latter may lead to different marks being awarded based on the weight given by different examiners to the contradictory instructions in the mark scheme.

One example where the mark scheme has not provided sufficient ways to determine a definitive mark for a mixed response is illustrated in the quote below. Here the examiner has made an overall evaluation of the response but is aware they are unable to match this to the articulation of the quality scale in the mark scheme. They commented on the different qualities and weaknesses of the response:

> See I'm looking at my comments and I think, I have done the balance. I've got paragraphs, question mark, but the language is good. And then I've got, really trying for content, but lapsed in control and cohesion. Some sentences are falling apart but they've got empathy. Quite a capable candidate but really uneven and a bit of a mixed bag. And that's what you've got here.

When this happens, examiners find it difficult to decide which band in the mark scheme is most appropriate for a response and this can lead to a range of marks being awarded. This issue of inconsistent performance across response features might sometimes be alleviated by splitting, for example, different assessment objectives into different columns on the mark scheme where different bands can be awarded to different features. However, we saw in the 'ignoring instructions around the independence of 2 marks' section above that this can sometimes have other issues. These kind of mixed responses are simply more demanding to mark than more even, consistent responses.

Exceptionalities

This sub-category of definitional uncertainty is when candidates' response is unusual or atypical, perhaps on a single quality dimension. For example, responses where the candidate can spell difficult words correctly but makes spelling mistakes on easy words. In most cases, the mark scheme does not cater for such cases and the bands simply go from simple spelling to complex spelling. When this happens, examiners

have to decide which band the candidate's response falls into and given there is no guidance on that specific instance, examiners may well come to a different conclusion and award different marks.

Oddities

This sub-category of definitional uncertainty is when a candidate includes something unusual in the response, an oddity, and it might influence different examiners in different ways. In a series of questions, the candidates had to write about an event they would organise for the school. Some examiners commented on the fact that the candidates' responses sometimes contained unrealistic claims. For example:

> - Yes, with £500 we can provide a small city with clean water. Well intentioned.
>
> - So how are you supposed to deal with something like that? Things that are slightly ridiculous.
>
> - I think it's a factor you say, […], they're going to have the most ridiculous people present. Rod Stewart is going to be there and etc. And you have to take that into account that this is totally unpersuasive and unrealistic. And you don't write them out because they do that, but it would influence the decision you make. It's got to.

This could lead some examiners to award lower marks if they chose to be influenced negatively on the basis of the oddity. Moreover, as the conversation above shows, it is unclear how to handle this phenomenon. This could lead to a range of marks being awarded. In this instance the task was to assess the quality of the writing and communication, and it was unclear whether this included plausibility of claims.

3.3 Observations made during the one-day meetings

There was an enormous amount of qualitative material gathered during the course of this research. Here are a few observations made during the one-day meetings:

1) Examiners were very keen to pursue the notion of the fairest mark. They are professionals trying their hardest to give the fairest marks. This is even more notable given that the marks discussed and argued about in this context would not actually be passed on to the candidates, but were essentially, an academic argument. Time pressures in live marking were often referred to in English language where examiners would like to be able to spend more time on each script but given the finite amount of time available, they feel they have to mark to the best of their ability whilst reading the response once. Annotations were often considered as important in this context as they are used as an aide memoire to for the whole response in order to come to a judgement after a first reading. Time pressured and single readings might have an impact on the ability to make inferences around 'subtle' responses ie those that contain some nuance which may not be immediately gleaned on a single reading.

2) When reviewing scripts, the mark awarded during live marking acted as an anchor for the examiners who, more often than not, tended to not move away from the mark awarded. For example, examiners marking the same responses blind would give

consistently lower or higher marks than examiners reviewing the response. This could be due to the available annotations during reviews of marking which made it obvious why and where the original examiner awarded the marks. Unless the examiner reviewing believed that the first marker had made an error, the reviewer tended to award the same mark as the first marker. Overall, there was less variation in the marks awarded during reviewing than in the ones awarded during the marking exercise.

3) In biology, examiners seemed not to look at an answer as a reflection of the candidate, whereas this seemed to be more the case in English language (see the section on "Some examiners making inferences on the basis of candidate characterisation"). This is probably due to the fact that biology has fairly short answers relating to a biology topic. In contrast, English language responses often provide a more personal perspective, and can sometimes include information that can give an indication as to the candidate's interests or attitudes. Moreover, it is possible that some English examiners might believe that, in general terms, 'writing is a reflection of the writer'. Examiners in English language sometimes made reference to candidates' attributes that they inferred from reading a response over a few pages. These inferences may have an impact on the marks awarded as they could act as a bias that will colour their view of a candidate's response.

4) We observed that the role of the PE seemed quite different in the 2 subjects. This was consistent across the different meetings within the 2 subjects. The role of the PE in English language units was more top-down, more hierarchical in style than biology. While in both subjects, PEs are at the top of the marking hierarchy, and therefore responsible for how marks should and should not be awarded, this role appears to have acquired a more special status in English language as a kind of 'ultimate arbiter'. This is probably because in English language, the open-ended nature of the responses and the levels of response mark schemes (ie the inherent definitional uncertainty present), mean that the PE, in order to 'set the marking standard', has to assume this greater role, and 'embody' the marking standard.

In terms of the dynamics in the meetings, researchers observed that the discussions in biology were focused on the correctness of biology present in the response and its relationship to relevant marking points in the mark scheme, arriving at a consensual view regarding the appropriate mark, or acknowledging that alternative views might be understandable. In biology, the senior examiners had a more equal status to the PE, who would listen and assimilate their views. In contrast, in English language, the discussions sometimes focused upon something more nebulous – the 'reading' of a response and its overall 'quality'. Senior examiners tended to be more deferential to the PEs, who, across all the English language meetings, had very much the final say on the mark, the 'correct' interpretation of the response and the mark scheme. One upshot of this special status appeared to be the openness with which senior examiners were willing/able to discuss different 'readings' of a response. While biology senior examiners were more than happy to discuss the rationale for their given mark, different from that of others (including the PE), English language senior examiners were frequently reluctant to disclose that they had given different marks unless they were in very close proximity (in the same level) to the PE's preferred mark.

5) In one of the biology meetings, the examiners had a discussion around specificity of language and questioned whether all candidates had a sufficient level of English and whether their achievement of marks was sometimes impeded by poor English skills rather than lack of biological knowledge.

6) Throughout the English language meetings, when examiners were looking at responses, a number of them made comments on the comparison between the current response and the last response they read (Vaughan, 1991). Examiners are usually told to mark the current response without casting their mind back to the previous answer for comparison. This seems to be a difficult instruction to follow and a number of examiners were comparing the qualities of the current and last responses, possibly anchoring their mark for the current response on the mark awarded to the previous one.

7) At times, it felt like examiners of the English language papers were "impression marking". Impression marking is a holistic marking method where there are no criteria to assess against. It might be that the examiners had internalised the mark scheme so well that they were able to maintain in their head the "standard" for each level of the mark scheme. However, it is possible that they first evaluate an answer using impression marking and then go back to the mark scheme to confirm their mark. This could explain why a number of examiners were using heuristics to award a mark.

3.4 Summary of results

Commonality between subjects

For both biology and English language, all responses were classified within the four existing categories. Both had examples of procedural errors where there was a clear misapplication of the mark scheme.

Differences between subjects

All the sub-categories, except 'clear misapplication of the mark scheme' were specific to biology or English language in the examples seen, though we suspect with a greater sample of work there would be more commonality. Mixed responses are present in both subjects but the sub-types are slightly different. The frequency of occurrence of the different types was also different between subjects with more inferential uncertainty instances in biology and more definitional uncertainty instances in English language. In English language, most instances of definitional uncertainty were sub-classified as mixed responses. The classification of responses in English language showed that inferential and definitional uncertainty were present in most responses. It can be quite difficult to disentangle inferential and definitional in mixed responses as they are sometimes present in isolation but the 2 types of uncertainty also sometimes interact.

Given there are a number of low tariff, short questions in biology, it appears that scanning and matching is a commonly used strategy that can lead to word spotting.

According to Black and Newton (2016), some instances of procedural errors occur when an examiner fails to look outside of the "clip[7]" area that is visible on the

[7] Clips are used when marking online and refer to the fact that the response presented shows the area comprised by the question and the answer space and not the whole page.

marking system. In this study, the bespoke marking system did not use clips but instead used whole scripts, making this sub-category of procedural error less likely.

Table 2 presents a list of the identified sub-categories for each subject.

Table 2. Table containing all the sub-categories identified in biology and English language.

	Biology	**English language**
Procedural error	Clear misapplication of the mark scheme	Clear misapplication of the mark scheme
		Applying the wrong generic mark scheme
		Ignoring instructions around the independence of 2 marks
Attentional error	Inattention resulting from word spotting	Inattention leading to missed information
Inferential uncertainty	Differing interpretations of handwriting – identification of words/letters	Heuristics based on characterisations of the candidate
	Differing interpretations of sentence/paragraph meaning	Heuristics based on superficial features
	Difference in making inferential leap (minor)	Subtlety
	Difference in making inferential leap (major)	Primacy/recency effect
Definitional uncertainty	Mixed response/differing decisions based on the mark scheme (correct and unrelated incorrect information)	Mixed response/differing decisions based on the mark scheme
	Mixed response – correct and incorrect/differing decisions based on the mark scheme (correct and contradicting incorrect information)	Exceptionality
		Oddities

4 Discussion

The results presented in this study support the original categories of marker disagreement proposed by Black and Newton (2016). Given this was the first systematic test of the typology, these are important results. The overall objective of the study is to help improve the quality of marking by increasing the reliability of marking. The further sub-categories identified in this study could be used to improve marking by either suggesting improvement to the mark scheme or providing more training or guidance on how to handle certain features. Further work in this area, analysis of more responses in a greater range of subjects, would be helpful to further

test the categories and sub-categories, and their prevalence across different subjects and different item and mark scheme types.

In coding some of the reasons for disagreement, there was, on occasion, 2 or more types occurring independently (in different places) in a response. However, we also found some interactions between types, particularly an interaction between inferential uncertainty and definitional uncertainty. It is possible that, for example, depending on the differing inferences made about a response, and on differing interpretations of the mark scheme, that a greater number (range) of marks are possible or justifiable.

A number of psychological forces have been identified as sub-categories in English language. For instance, given the complexity of marking, some examiners may use heuristics in order to alleviate inferential uncertainty, which may lead to different marks being awarded. More specifically, the use of the response's length (Hall and Daglish, 1982), the quality of handwriting (Briggs, 1980), primacy/recency effect, topic, implied candidates' characteristics (Baird, 1988) and others can all be used as shortcuts when it comes to deciding on the mark to award. Heuristics are generally automatic and unconscious but it is also possible that information on the different cognitive processes involved in the heuristics could help reduce their use and hence reduce marker disagreement. In some of the instances in this research, it appears that such heuristics can lead to marks which are not fair representations of the quality of the response and exam boards are likely to wish to seek ways to discourage this. These heuristics or biases have been extensively studied in marking (see Meadows and Billington (2005) for a review). This study did not set out to study these biases directly but it has identified their importance in marker disagreement. Most of these biases do not have the same impact on all examiners and this is why they award different marks. The fact that Black and Newton's framework (2016) is able to accommodate these biases/heuristics is an important feature.

Another way of reducing marker disagreement might be to encourage candidates to pay more attention to the way they construct their answers. The use of short sentences with punctuation, logical structure and good quality of English could help reduce some instances of attentional errors and inferential uncertainty.

4.1 Limitations

A few limitations have been identified in the study. First, the examiners in the meeting did not necessarily know or remember what mark they had awarded for each question so their comments on why there are discrepancies may have sometimes been hypothetical or else trying to reconstruct their rationale subsequently. Without a more 'immediate' method, such as a verbal protocol ('think aloud') study, we cannot be certain that what they suggested in terms of the process or reasoning of awarding the mark was that which actually happened. As discussed earlier, the fact that there was no annotation facility during the marking also prevented the examiners from being reminded where they had awarded marks by looking at their annotations. The lack of annotation facility may have had a larger impact in English language as the subject includes only extended responses.

When carrying out the one-day meetings, the responses were printed and examiners were given as much time as needed to read them again during the meeting. This way of marking was different than how the scripts were marked during the marking study itself and examiners may have taken more time to read the responses and used their pens to make annotations, leading to different determinations for the mark.

Finally, the scripts used in the study were by no means representative of the scripts that are marked during the live marking window as they were taken from scripts sent to be reviewed (reviews of marking) and therefore are likely to, as a set, represent those responses which are harder to mark.

4.2 Conclusion

Overall, this study lends support to Black and Newton's (2016) categories of error and uncertainty. Moreover, it has started to define or exemplify the categories further, using responses in biology and English language.

The ultimate aim of this type of study is to understand the complexity of marking so ways to improve the reliability of marking might be identified. How could this study help reduce examiners' disagreement? Some of the sub-categories we identified could be handled by having more guidance or training. For example, more guidance should be available to examiners on how to handle mixed responses that contain both correct and incorrect information so that all examiners can adopt the same strategy. Also, some of the procedural errors could be reduced by using less generic mark schemes and by providing more guidance and training on how to mark responses that have different qualities on 2 different assessment objectives. Finally, guidance to candidates on how to construct their answers logically and grammatically could help reduce attentional and inferential uncertainty.

Given the results in both subject were different, it would be interesting to replicate this study with other subjects to see whether more sub-categories could be identified. Also, it would be interesting to see whether modifying the mark scheme to include more information on how to handle mixed responses could reduce definitional uncertainty.

References

Aldrovandi, S., Poirier, M., Kusev, P. and Ayton, P. (2015) *Retrospective evaluations of sequences: Testing the predictions of a memory-based analysis. Experimental Psychology, 62 (5)*, 320-334.

Baird, J. (1988) What's in a name? Experiments with blind marking in A-level examinations. *Educational Research, 40,* (2), 191-202.

Black, B and Newton, P. (2016). Tolerating difference of opinion. Paper presented at the 17th Annual AEA-Europe Conference, Cyprus.

Black, B., Suto, W.M.I. and Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy and Practice, 18*, 295-318.

Briggs, D. (1980) A study of the influence of handwriting upon grades using examination scripts. *Educational Review, 32* (2), 185-193

Hall, C. G. & Daglish N. D. (1982) Length and quality: An exploratory study of inter-marker reliability. *Assessment & Evaluation in Higher Education, 7* (2), 186-191.

Kahneman, D., and S. Frederick, (2002). Representativeness revisited: attribute substitution in intuitive judgement. In Heuristics and biases: the psychology of intuitive judgement, ed. T. Gilovich, D. Griffin,

Stanovich, K.E., and R.F. West. (2002). *Individual differences in reasoning. In Heuristics and biases: the psychology of intuitive judgement*, ed. T. Gilovich, D. Griffin and D. Kahneman, 421–440. Cambridge, Cambridge University Press.

Suto, W.M.I., and J. Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal 34 (2*), 167–187.

Vaughan, C. (1991) Holistic assessment: What goes on in the rater's mind? In L. H.-. Lyons (Ed.), Assessing Second Language Writing in Academic Contexts. Norwood, N.J.: Ablex Publishing Corporation.

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

November 2018 **Ofqual/18/6449/5**