# Research into alternative marking review processes for exams

December 2015

Ofqual/15/5804

# Contents

# 1. Executive summary

This research looked at four different potential processes for enquiries about results (EAR) for examined work:

- The **current review process** – in which examiners review the original marking of the script and should change the mark only where the original mark cannot be justified. (condition 1)

- A **review plus tolerance** process – in which examiners review the original marking as above. A 'tolerance' is applied such that any second mark within tolerance (say, ±3 marks) of the original mark is discarded and the original mark retained. This would be to prevent one legitimate mark being replaced by another legitimate mark. (condition 2)

- A **single clean re-mark** – where examiners mark the script with no sight of the original marking or annotations. This is sometimes called a blind re-mark. (condition 3)

- A **double clean re-mark with resolution** – where two examiners independently mark a script. Where their marks differ, they come to a joint decision on the most appropriate mark. (condition 4)

Three exam units from three boards were used in this study, with 32 examiners in total. The scripts had all been EAR scripts in the summer 2014 session and all the examiners had also been EAR examiners for that session.

The findings from this research are summarised as follows:

- The mark changes in the study were generally small (on average, less than 1 mark). The majority of mark changes (more than 85 per cent) in all units in all conditions were within 3 marks.

- The condition that is least likely to result in a mark or grade change is review plus tolerance (condition 2). However, we also looked at the accuracy of marks by estimating, for each script a 'true score'. The true score is the notional score of a candidate were there no random error in the measurement. Comparing the marks in condition 2 with an estimate of true score[1] indicates that imposing a tolerance introduces error – probably because it suppresses a marker's expert

[1] In this research an estimate of true score was generated by calculating for each candidate script the mean mark produced by the independent clean marks gathered in condition 3. All references to true score are a shorthand for estimated true score.

judgement. In other words, tolerance would not be a fair method of preventing the substituting of one legitimate mark for another.

- The condition that produces outcomes closest to true score is double clean re-mark with resolution (condition 4). The margin between this condition and the next best – the current review process (condition 1) – is around 5 to 6 per cent of scripts for up to within 4 marks' proximity to true score. This margin is noticeable but not substantial given the potential costs and difficulties of introducing such a model.

- The current review process (condition 1) performed well in terms of proximity to the true score, second only to double clean re-mark with resolution.

- Single clean re-mark (condition 3) involved greater mark variability and did not perform as well as some other conditions in terms of proximity to the true score. It appears that the apparent advantage of markers being unbiased by the removal of the original marking might be more than offset by the risk of introducing some (new) small errors. Examiners themselves also rated this method as least likely to deliver fair and accurate results.

- Data from the questionnaire indicates that for the current EAR process (review), some examiners are adopting slightly different processes. Some examiners indicated that they were reviewing the marking of another examiner, while others indicated that they adopted some elements of the re-mark process. Greater guidance and training is probably needed in order to ensure greater consistency of approach.

# 2. Background

After GCSE and AS and A level results have been issued, it is possible for candidates who believe they have received the wrong mark or grade to have the original marking checked. This process is currently called enquiries about results (EAR). Of the three EAR services, this research looks at service 2 – post-results review of marking.[2] The current check involves a review of marking whereby the reviewing examiner can see the original marking and annotation. While the provision of these services and deadlines are currently laid out in our *GCSE, GCE, Principal Learning and Project Code of Practice*,[3] which is due to be withdrawn, some of the details are also managed by the Joint Council of Qualifications to help ensure that its members offering GCSEs, AS and A levels follow a common approach.

As the title of service 2 – post-results review of marking – suggests, this is not a clean or 'blind' re-mark where any script is marked afresh with no visibility of the original marking. Instead, it is a "review of the original marking to ensure that the agreed mark scheme has been applied correctly" (JCQ, 2015). There are two distinct key concerns frequently aired around how this process within service 2 operates. First, there is a perception that this review approach tends to be overly confirmatory – a 'rubber stamping' approach (for example, HMC, 2012) – the perception that incorrect marking is not remedied or redressed. The concern is that by seeing the original marking, examiners are 'biased' or overly inclined to award the same mark as originally issued.

The second concern, paradoxically, is that there are too many changes. Many changes in the EAR process are small changes of 1 or 2 marks, even on essay-based subjects. Given that there is a small but acceptable amount of variability in marker judgement, small mark changes in the EAR process suggest that reviewing examiners are simply replacing one legitimate mark for another legitimate mark. This is of concern because it suggests that the system might be allowing candidates who received a legitimate mark in the first place, particularly those very close to qualification grade boundaries, an avenue to enhance their overall grade and this could be perceived as unfair to those candidates who did not submit an EAR.

The following section looks at the extent to which there is evidence for these two claims.

[2] There are two types of service 2 – priority (available for AS and A level) and non-priority (available for GCSE and AS and A level), the former having tighter deadlines. The other two EAR services are clerical check (service 1) and post-results review of moderation (service 3) for internally marked and externally moderated units.

[3] www.gov.uk/government/uploads/system/uploads/attachment_data/file/371268/2011-05-27-code-of-practice.pdf

## Previous research evidence on blind versus non-blind re-marking

There is evidence from marking studies that when the second examiner can see the original marking and annotations, there is greater mark agreement than when the marks are removed.

Murphy (1979) sent previously marked scripts from an unnamed exam to two senior examiners – 100 with original marks and comments and 100 clean – and found a considerable difference in rates of marker agreement. Visibility of original marks gave a mark/re-mark correlation of 0.94, compared to 0.87 for blind re-marking. The mark differences tended to be smaller in the sighted re-mark, with a mean absolute mark difference of 2.74 compared to 5.51 for blind re-marking. This indicates that the second mark is not truly independent of the first unless the original marks are removed.

Vidal Rodeiro (2007) found a similar pattern looking at English and classical Greek GCSE units. Higher correlations were found for sighted re-marks compared to blind re-marking treatments, and again smaller mean absolute mark differences were found – 0.67 for sighted re-mark versus 1.97 or 2.16 for blind (depending on the seniority of the markers). Again, this indicates that visibility of original marking has some kind of anchoring effect.

Similarly, Meadows and Baird (2005) looking at standardisation scripts, found that when senior examiners marked clean (photocopied) scripts compared to annotated (live) scripts, there were greater mark differences in the blind marking condition.

Investigating possible systems for double marking, Fearnley (2005) used two subjects (English and business) and found that double marking systems using cleaned scripts rather than annotated scripts produced a small increase in marking reliability – the implication being that where a second marker could see the first set of marks, their own marking was unduly influenced and insufficiently independent.

Billington (2012) looked at monitoring methods of examiners in live marking, comparing two scenarios: (1) where senior examiners monitored assistant examiner live marking with the original examiners’ marks visible; and (2) where assistant examiners’ were given definitively pre-marked and then cleaned scripts. The former scenario appeared to overestimate the accuracy of assistant examiner marking.  This was also a finding of Meadows and Baird (2005).

To summarise the research on blind versus non-blind re-marking, it seems clear that non-blind re-marking tends to provide closer agreement to the original mark and this could be seen as a source of bias in the marking of the second examiner. However, the context or aims of all of this research is slightly different from that of the current question. While the current question is around the most appropriate method for conducting an EAR, the research above has been undertaken in different contexts –

standardising markers, monitoring markers or marking reliability studies more generally. Therefore, the extent to which the same findings would apply to markers knowingly reviewing scripts of candidates in an EAR is not clear. The research base is also less clear on which method produces marks most closely associated with the 'true' mark or score[4] – undoubtedly a key question for the context in hand whereby an EAR should primarily be about remedying error for candidates. It is at least hypothetically possible that a single blind re-mark may not offer any advantage over reviewing in terms of proximity to true score.

## Evidence around small mark changes in EAR

Some small mark changes in Enquiries about Results reflect the correction of genuine marking error. But some small mark changes may reflect acceptable differences of academic opinion between two equally skilled professionals. In normal live marking processes, exam boards use the concept of 'tolerance'[5] to denote the level of acceptable mark variation that could be given to the same answer or script.

In 2014, over 80 per cent of mark changes were small mark changes – changes within the original marking tolerance.[6] This was in line with the 78 per cent of mark changes within the original marking tolerance that we reported for 2012 data in our report on quality of marking and which included qualifications equivalent to GCSE and A level.[7] In this report, we looked at two subjects – geography and French – in more detail and the figures are reproduced here (see figures 1 and 2).

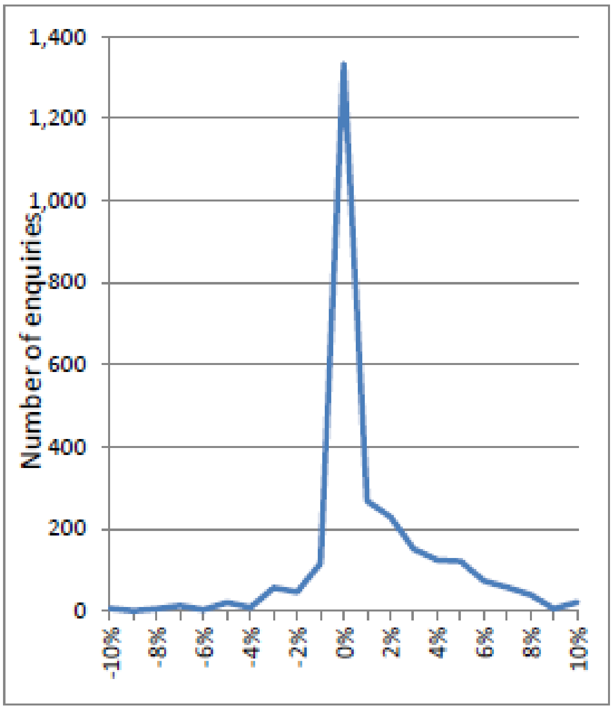
Both subjects showed low rates of large mark changes – less than 1 per cent of mark changes were 10 per cent or more of the 'raw' marks available for the unit. Most significantly, around 60 per cent of qualification grade changes for both subjects were within the original marking tolerance. This data appears to suggest that the majority

[4] In classical test theory, a true score is the notional score of a candidate where there is no error in the measurement. The true score is defined as the mean score of an infinite number of observed scores (marks) in independent administrations of the test. The best approximation to a true score from one particular administration of a test can be obtained from taking the mean from multiple independent measurements of the work. In the current research, we were able to derive a true score for each script because each script was marked multiple times in condition 3 by independent examiners. We could then compare all script marks to the true score.

[5] The marking tolerance is a measure used during live marking to judge whether an examiner's marking is acceptable. It varies according to the subject and type of question, but generally reflects the legitimate difference of opinion between two equally skilled examiners. See *Delivery of Summer 2014 General Qualifications*: www.gov.uk/government/uploads/system/uploads/attachment_data/file/386477/Delivery_of_summer_2014_general_qualifications.pdf

[7] *Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications*: www.gov.uk/government/uploads/system/uploads/attachment_data/file/393832/2014-02-14-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-final-report.pdf

of mark and grade changes are the result of substituting one legitimate mark for another legitimate mark.

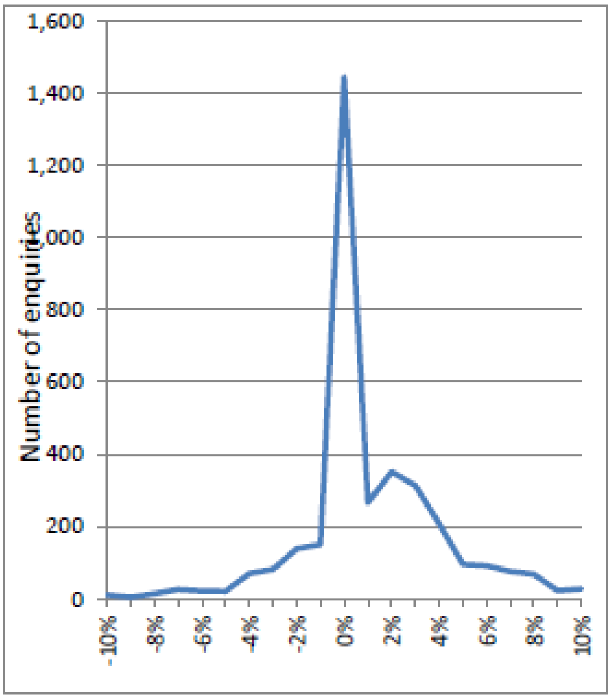


**Geography A level and equivalent**

- 4,307 enquiries about results in summer 2012 (3 per cent of all scripts);
- forty per cent of enquiries about results led to no mark change;
- eighty-seven per cent of mark changes were within 5 per cent of the total raw marks available for the unit;
- one per cent of mark changes were 10 per cent or more of the total raw marks available for the unit;
- 769 qualification grade changes were made;
- sixty-six per cent of these were the result of mark changes within the original marking tolerance.

**Figure 1. Mark differences in EAR as a proportion of mark scale for geography A level and equivalent**[8]

[8] Source: *Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications*: www.gov.uk/government/uploads/system/uploads/attachment_data/file/393832/2014-02-14-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-final-report.pdf



**French A level and equivalent**

- 3,270 enquiries about results in summer 2012 (5 per cent of all scripts);
- forty-nine per cent of enquiries about results led to no mark change;
- ninety-one per cent of mark changes were within 5 per cent of the total raw marks available for the unit;
- one per cent of mark changes were 10 per cent or more of the total raw marks available for the unit;
- 428 qualification grade changes were made;
- fifty-nine per cent of these were the result of mark changes within the original marking tolerance.

**Figure 2. Mark differences in EAR as a proportion of mark scale for French A level and equivalent**[9]

The majority of scripts submitted for EAR are from candidates who are very close at overall qualification level[10] to the next grade boundary.[11] The majority of candidates are a few uniform mark scale marks (UMS) under the next qualification-level grade boundary and this may represent only 1 or 2 raw marks. Research conducted on our behalf suggests that some schools and colleges are aware of a level of subjectivity in the system and how the EAR process can work. Some teachers referred to entering EAR just below grade boundaries as a “one way bet”,[12] given the very low likelihood of the grade going down but reasonable likelihood of it going up.

If the current process is allowing one legitimate mark to be substituted for another, it becomes an issue of equity – those who can and are prepared to pay for an EAR

[9] Source: *Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications*: www.gov.uk/government/uploads/system/uploads/attachment_data/file/393832/2014-02-14-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-final-report.pdf

[10] While EAR scripts are marked at unit level, looking at the effects at qualification level – that is, for the candidate’s overall GCSE or AS/A level grade – is most relevant.

[11] *Delivery of Summer 2014 General Qualifications*: www.gov.uk/government/uploads/system/uploads/attachment_data/file/386477/Delivery_of_summer_2014_general_qualifications.pdf

[12] Oxygen, 2014, p12, quoted in *Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications*: www.gov.uk/government/uploads/system/uploads/attachment_data/file/393832/2014-02-14-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-final-report.pdf

might be advantaged in receiving another legitimate mark that they prefer, while other candidates who cannot or who are not prepared to pay do not have access to this same possibility.

## Researching the options

Given the issue of equity, one possible change considered to the EAR process was for exam boards to use a tolerance, much as is currently operated in live marking. This would mean that a review mark within the tolerance of the original mark would not be passed to the candidate on the basis that it would just be substituting one legitimate mark for another legitimate mark. This model would prevent the scenario where one legitimate mark is substituted for another legitimate mark. Where an EAR review indicates a small mark change within the original live marking tolerance, in this model, the newer mark would not be passed on to the candidate – their mark would remain at the original mark. This model would not of course stop EAR scripts with large mark changes or definite marking errors (no matter how small) being remedied. It would only prevent small judgemental mark changes within tolerance being passed on to candidates. There are variations on this proposal in terms of how marks outside of tolerance can be treated – for example, they might automatically replace the original mark, or they might be subject to additional escalation where a third but blind re-mark determines which of the original mark or second mark is most legitimate. Overall, the key hypothetical advantage of this review plus tolerance approach would be to directly prevent substituting one legitimate mark for another.

Another possible approach is that of a single clean re-mark (sometimes called a blind re-mark). This approach has sometimes been suggested as a more desirable alternative for the EAR process on the basis that examiners would not be guided or biased by the original marking, as the research mentioned above indicates. However, there are questions around the appropriateness of a single blind re-mark in the context of an EAR – for example, to what extent might such an approach give a more accurate result (meaning equal or close to the true score). One potential risk of a single clean re-mark is that the second examiner makes a mistake – such as misreading part of an answer, misinterpreting part of an answer, or missing part of an answer – which may or may not be different from any original marking error. The argument here is that a blind re-marker may not improve the accuracy of marking because they cannot benefit from the original thinking presented in the original annotation. In a sense, a review of marking could be conceived of as a non-blind double mark. Fearnley (2005) found that such a process could be advantageous over a single examiner mark.

Because of the risks associated with a single examiner clean re-mark, we decided to also compare this to a condition in which each script is marked by two independent examiners with no sight of the original marks or annotations. This should mitigate risks of a single examiner introducing an error. This double marking model was found to provide significantly better agreement with the definitive marks in Fearnley's (2005) study in both subjects, although only by a small margin (less than 1 mark on average).

## This research

The primary aim of this research was to determine the impact of different EAR processes on the final outcomes in terms of (1) rates of mark and grade change and (2) proximity to the true score as a way of evaluating the accuracy of any approach overall.

This research will help us to evaluate the implications and appropriateness of any change to the procedure for dealing with written exams through the EAR process.

# 3. Methodology and design

## Conditions

We investigated four possible models or research conditions for the review of marking to understand the impact of each on the amount and nature of mark changes.

**Condition 1 – in which markers reviewed the original mark in line with their understanding of the current process.** In the current scheme, reviewing markers can see the original marks and annotations and should only change marks where the original marker misapplied the mark scheme. In this study, markers were reissued with their standard instructions for conducting EAR. For shorthand, we will call this research condition the **current review process**.

**Condition 2 – in which the marker reviewed the original mark, but that mark was only changed if it was 'outside tolerance' and not changed if it was 'within tolerance'.** This model was included as it was considered a possible means to prevent one legitimate mark being replaced by a different legitimate mark. The tolerance would take the form of a predetermined mark range and the original mark would not be changed if the reviewing marker's alternative mark was within that range from the original mark. In this research, we modelled the effects of tolerances of just ±2 and ±3 marks – equivalent to between ±2 per cent and ±3.75 per cent of raw marks depending on the maximum mark of the question paper. In this research, markers reviewed the scripts and submitted marks knowing that a tolerance would be applied. For shorthand, this research condition is called **review plus tolerance**.

**Condition 3 – in which the marker marked the assessment afresh without seeing the original mark or any comments made by the original marker.** This model was included because some critics of the current system suggest that reviewing markers are unduly influenced by the original marker's mark and annotations. They argue that the second marker should mark a clean copy of the script and so be uninfluenced by the views of the first marker. Because the marker is unaware of the original mark, for shorthand, this condition will be referred to as **single clean re-mark**.

**Condition 4 – in which two markers independently marked the assessment afresh without seeing the original mark or any comments made by the original marker. If the marks do not agree, the two markers come to a joint resolution on the final mark through discussion.** In some ways, this might be considered the 'Rolls Royce model' in that it satisfies some people's desire for a blind or clean re-mark without introducing any risk – for example, of a new clerical error. For shorthand, this will be referred to as **double clean re-mark with resolution**.

## Selection of units

It was important that the study involved more than one unit and should reflect some of the diversity of item types and exam styles and levels in the GCSE and AS/A level system.

Three awarding bodies in England each provided one unit for the study. Suitable units were selected on the following basis:

- A large entry or volume of EAR – this gives us a reasonable number of EAR examiners to take part in the study and helps us to generalise our findings.
- A wide range of subject areas – between them, the three units reflect a range of subject areas (including a humanities and a science subject).
- A range of profiles of item types – each unit has a different profile of item types (as summarised in table 1). It is known that different types of item are associated with different levels of marker agreement (see, for example, Black, Suto and Bramley, 2011). This helps us to understand the extent to which different possible EAR processes work in different types of exam with different profiles of item types.

**Table 1. Summary of units according to profile of item types and mark schemes**

| Unit / Board | Item type | Mark scheme type |
|---|---|---|
| A | Large number of items – a mixture of objective items, short-answer questions and extended response | Objective- and points-based |
| B | Large number of items – predominantly short-answer questions with low mark tariffs | Points-based |
| C | Small number of items – all medium response or extended response items | Levels-based |

## Examiners in the study

In total, 32 examiners participated in the research. All the examiners were senior examiners for those units and had marked EAR scripts during the 2014 session for one of the units in the study. For one unit, the examiners in the study represented the entire EAR panel of examiners. For two units, the examiners were broadly representative of the EAR examiners in terms of role seniority (team leader, senior team leader, assistant principal examiner, principal examiner).

**Table 2. Summary of scripts and markers**

| Unit / Board | Markers | Unique scripts in study | Total number of scripts marked per marker |
|---|---|---|---|
| A | 12 | 120 | 160 |
| B | 8 | 80 | 160 |
| C | 12 | 80 | 160 |

## Sampling of scripts

All of the scripts in the study had been through the EAR process in 2014. They were sampled to make sure that they were representative of each unit's EAR scripts in terms of:

- Original raw mark
- Mark (and UMS) change at EAR
- Original qualification UMS – for each candidate, this is their overall UMS when the UMS for all units within the qualification are added together.

A comparison of all EAR scripts and the scripts used in the study are summarised in Appendix A.

The qualification UMS distribution was closely matched where possible. This is because for all the units in the study, the qualification UMS distribution was not a normal distribution but had multiple peaks approaching the location of grade boundaries. This pattern of steep cliffs and drop-offs around grade boundaries has been noted previously across a number of units.[13] It is significant in that a small mark change can bring about a grade change at qualification level. For example, a raw mark change of 1 or 2 marks will mean a similarly small change in UMS, but it can be sufficient to take the overall UMS across the threshold from say 179 UMS (equivalent to a grade D at A level) to 180+ UMS (equivalent to grade C).

Figure 3 illustrates this pattern of peaks and drops for one AS unit in the study. As far as possible, the selection of scripts aimed to replicate the pattern of proximity to qualification UMS grade boundaries. This was to help to accurately gauge the impact of any mark (and therefore UMS) change on qualification grade outcome.

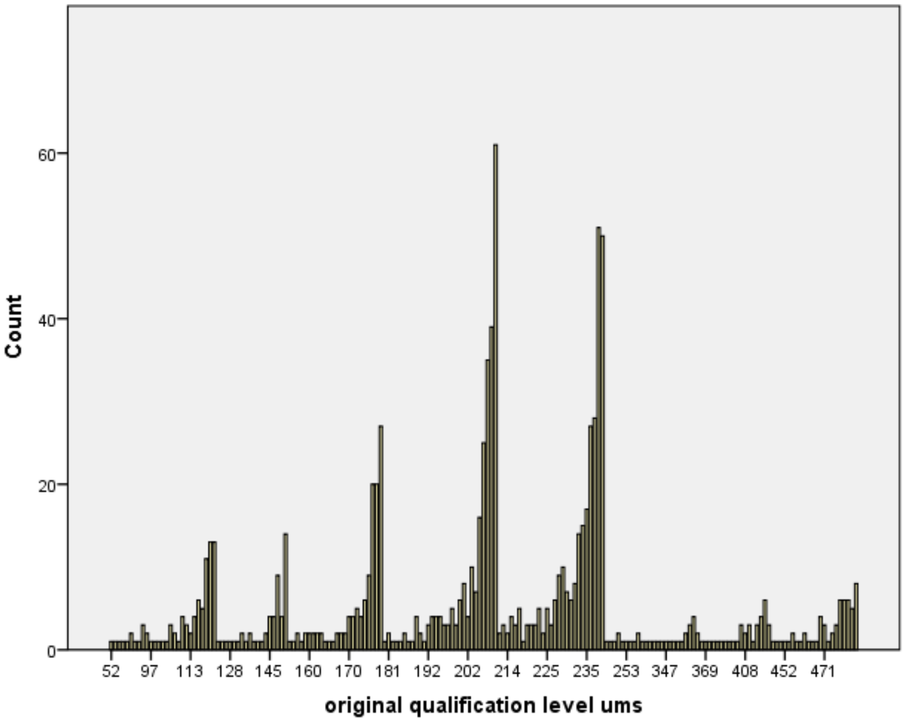


**Figure 3a. All EAR scripts in 2014 for one unit, by original qualification-level UMS**

[13] *Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications*: www.gov.uk/government/uploads/system/uploads/attachment_data/file/393832/2014-02-14-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-final-report.pdf

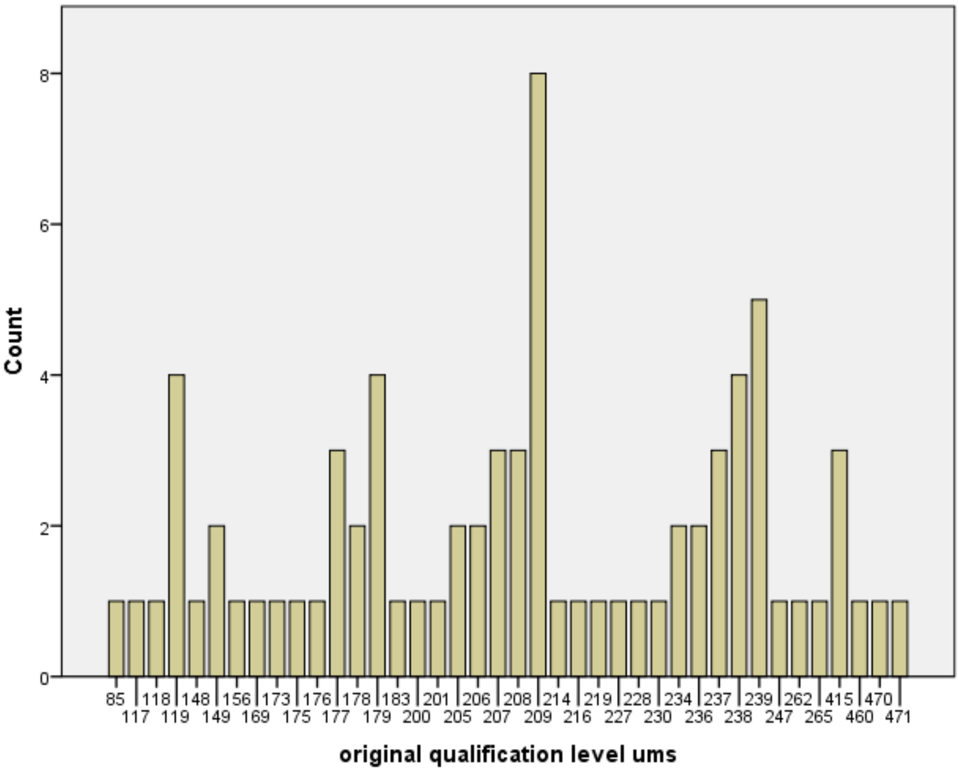


**Figure 3b. Sample of EAR scripts selected for research for the same unit as figure 3a, by original qualification-level UMS**

## Procedure

Scripts were received from exam boards in electronic PDF format. The scripts had been prepared to ensure anonymity of candidates.

Scripts were dispatched as hard copy to examiners. While all the units in this study were marked online (both for prime marking and EAR), it was not possible for exam board software to be used in a research capacity and other potential online marking solutions did not easily support multiple marking of scripts.

The first task for examiners to complete was to re-familiarise themselves with the question paper and mark scheme for their unit and go through a re-standardisation process in order to qualify to mark in the study. Each examiner had to mark five scripts and submit their marks to us. Feedback on the accuracy of the marks was given at item level and examiners' marks had to be in normal marking tolerances in order to continue in the study. A further five scripts were available should an examiner fail the first set of five, but no examiner did fail.

Condition 1 (current review process) and condition 2 (review plus tolerance) were marked first. The conditions were counterbalanced so that within each unit, half the examiners marked according to condition 1 first and half according to condition 2 first.

Controls were in place in terms of script order, so that the order of scripts marked was unique for each examiner and so that it was not the case that any particular script or set of scripts was marked by all examiners first or last.

For conditions 1 and 2, examiners were instructed to follow their normal instructions as issued by their board. These were dispatched with the scripts.

For condition 2 (review plus tolerance), examiners were told that any rules around tolerance would be applied by the researchers. The reasons for this were twofold. First, if tolerance were to be used in the live process, it is more than likely that boards would choose to apply it themselves rather than for examiners to apply it on the grounds that this would give the clearest audit trail. Second, for the purpose of research, we wanted to be able to model effects of different tolerance, so we needed to have a copy of the mark before tolerance was applied.

Essentially, this means that for examiners, the process for condition 2 should have been identical in process to condition 1 but for the fact that examiners knew that subsequently tolerance would be applied.

Once all the scripts in conditions 1 and 2 had been marked, they were returned to us and no record of the marks was retained by examiners.

Scripts for condition 3 (single clean re-mark) were dispatched. All of these scripts had had marks and annotations from the original marking removed, so examiners were marking these scripts fresh. While examiners would have been exposed to some or all of these scripts in condition 1 or 2, it was thought that given the volume of scripts, it would be unlikely for examiners to be able to recall the marks awarded previously. For condition 3, examiners had to submit an electronic record of the marks as they went along and retain the scripts.

Condition 4 (double clean re-mark with resolution) used the marks from the scripts generated in condition 3. Each examiner was pre-paired with another examiner for one or two sets of scripts for all 80 scripts marked in condition 3. Examiners received the contact details (email and phone number) of their paired examiner for a set of scripts, as well as a file showing the two sets of examiners' marks for each script, highlighting which item marks did not agree. Examiners were instructed to arrange to call one another to jointly discuss where their marks did not agree and come to a joint decision on the most appropriate mark. They had to record their decisions on the scripts (in a different colour pen) as well as on the electronic file. After the marking activities, scripts and files were returned to us.

Finally, examiners completed an online questionnaire that covered aspects of their normal EAR practices as well as the different processes in the research.

# 4. Results

Analysis of results included looking at rates of mark and grade change as well as the proximity to the true score.

## Mark and grade change for different models of EAR

Results are first described for one unit in the study – unit A – and then for all units overall to help build up a picture of the results. Table 3 gives a summary of mark changes for unit A to understand the overall profile of mark change according to condition, and allowing comparison with how the same script marks were changed in the actual live EAR process in 2014.

**Table 3. Summary of mark changes (unit A)**

| Condition | | Largest negative mark change | Largest positive mark change | Mean mark change (later mark − original mark) | SD |
|---|---|---|---|---|---|
| Baseline | Actual EAR process | −13 | 10 | 0.39 | 2.27 |
| Condition 1 | Current review process | −14 | 9 | −0.24 | 2.66 |
| Condition 2 | Review plus tolerance (marks as submitted before tolerance applied) | −18 | 7 | −0.31 | 2.78 |
| Condition 3 | Single clean re-mark | −18 | 12 | −0.53 | 3.59 |
| Condition 4 | Double clean re-mark with resolution | −18 | 9 | −0.10 | 3.42 |

The other two units show a very similar pattern (see Appendix A). Table 3 shows a comparison of the original mark with the new mark overall for each research marking condition. Because all the scripts had been through the actual EAR process in autumn 2014, we can also compare each of the research conditions with the actual

EAR process as a baseline. In this sense, we can regard this as a baseline to help understand how authentically the markers in the study behaved.

The key points from table 3 are summarised as follows:

- Condition 1 (current review process) – we might expect that this condition and the baseline actual EAR data would look quite similar given that the same review process was utilised in both. They are similar in most respects, with one exception: the mean mark change is positive (+0.39) in the live process, but negative (−0.24) in the research version. This means that mark changes were more likely to be negative in the research process in this unit. Other units also show a lower mean mark change (although not negative) compared to the live EAR (see Appendix A).

- Condition 2 (review plus tolerance) – the table shows only the marks as submitted, before any tolerance was applied. These are very similar to condition 1 (current review process), as we might expect given that the process for examiners in both conditions is identical.

- Condition 3 (single clean re-mark) – this condition has produced the most negative mean mark change as well as the largest standard deviation. This indicates that this condition produced the biggest changes and that these changes were more likely to be negative when compared to the live and/or current process. This indicates that this condition is likely to give the most variability in mark changes.

- Condition 4 (double clean re-mark with resolution) – this condition produced a slightly less negative mean mark change and less mark variability than condition 3 (single clean re-mark).

Table 4 gives a picture of whether different marking conditions are more or less likely to give mark changes and whether these mark changes are more or less likely to be positive or negative in Unit A (mark change data can be found in Appendix B).

**Table 4. Percentages of mark changes (unit A)**

| Condition | | % scripts with negative mark change | | % no change | % positive change |
|---|---|---|---|---|---|
| Baseline | Actual EAR process | 23.3 | | 35.8 | 40.9 |
| Condition 1 | Current review process | 37.6 | | 25.1 | 37.3 |
| Condition 2 | Review plus tolerance | No tolerance – marks as submitted | 42.1 | 20.4 | 37.5 |
| | | Tolerance = 2 marks | 15.2 | 74.0 | 10.8 |
| | | Tolerance = 3 marks | 8.3 | 86.7 | 5.0 |
| Condition 3 | Single clean re-mark | 46.6 | | 14.3 | 39.1 |
| Condition 4 | Double clean re-mark with resolution | 42.1 | | 15.6 | 42.3 |

The key points from table 4 are summarised as follows:

- Once again, when we compare condition 1 (current review process) with the baseline actual EAR data, we might expect these to look similar because the examiners are given the same task of reviewing scripts. While the proportion of scripts with mark increases in both the baseline actual EAR scenario and the research scenario are very similar (40.9 per cent and 37.3 per cent, respectively), the proportion of scripts with no mark changes is lower (25.1 per cent, compared to 35.8 per cent). Meanwhile, scripts with downward mark changes were more common in the research condition than in the live condition (37.6 per cent and 23.3 per cent, respectively). This indicates that markers in research were less likely to retain the original mark and more likely to reduce marks. This was true for all three units in the study. We can only speculate why this difference is present in the research. There are two main hypotheses. One is that markers in research, some months later and marking on paper rather than on screen, became slightly more severe in the research generally (although this is not reflected in the upward mark figures). Alternatively, in the research scenario, markers are more inclined to remove marks where it is justified because they know this cannot impact upon the actual candidate. It is possible, then, in the live scenario, that markers may feel reluctant to remove marks because they are mindful of the potential negative impact on the candidate. This possible interpretation reminds us that markers undertaking EAR may be making their judgements in the context of being aware of the possible ‘plight’ of the candidate whose script they are reviewing or re-marking. This may in some cases exert some influence on their decisions whether or not to remove a mark.

- Single clean re-mark (condition 3) was least likely to result in no mark change and most likely to result in a downward mark change.
- Applying the tolerance, as expected, increases the likelihood of no mark change. In this unit, even a 2-mark or 3-mark tolerance makes a substantial difference to mark changes compared to the marks as submitted (from 20.4 per cent for marks as submitted, to 74 per cent for a 2-mark tolerance and 86.7 per cent for a 3-mark tolerance). This indicates, for this unit, how a high proportion of altered marks are within a very small mark range of the original mark. Very similar results were found for the other two units.

What does each of the models mean for grade change rates at qualification level? Not all mark changes will mean a grade change, but because of the proximity of many candidates to the overall grade boundary, even just 1 or 2 marks might result in an amended grade.

Figure 4 shows the proportion of grades staying the same or moving up or down for each condition for unit A in the study. Condition 2 (review plus tolerance) is most likely to give no grade change, particularly when the larger tolerance of 3 marks is applied. In this unit, conditions 3 and 4 were slightly more likely to result in upward grade movement compared to the current review process. However, this was not the case for the other units in the research (see Appendix C). Figure 5 shows the overall qualification grade change where conditions 3 and 4 produce similar upward grade change rates to condition 1.

So, if it were desirable to have a process that is least likely to result in grade change as a result of small mark changes, review plus tolerance would be the obvious choice. However, we need to evaluate each of the models against the likelihood of the model producing a mark that is accurate or true.

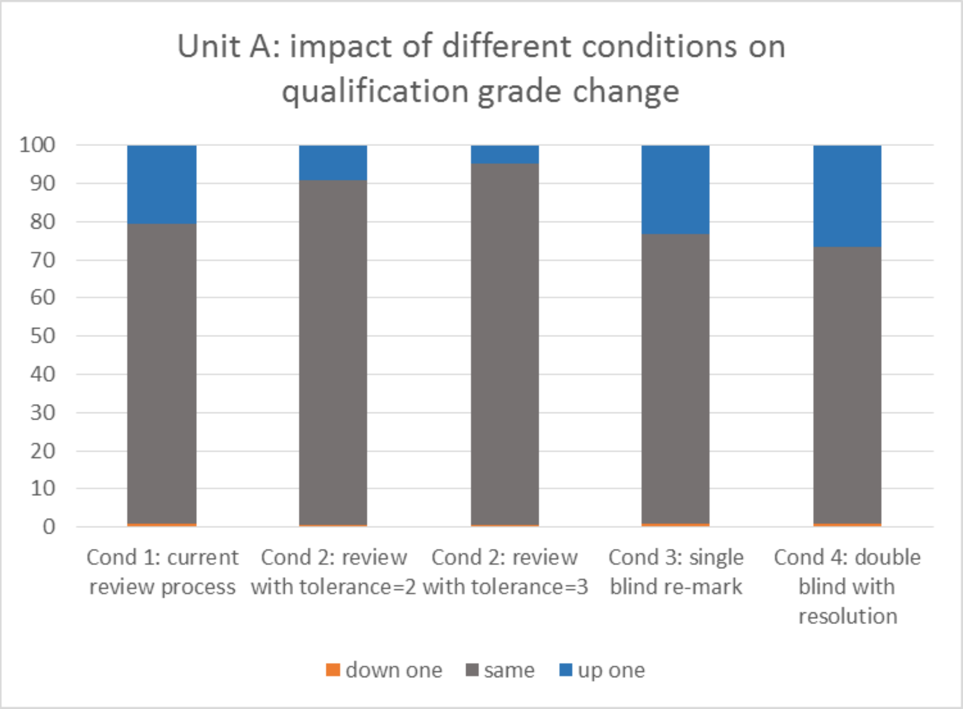


**Figure 4. Qualification-level grade change modelled according to different conditions (unit A)**

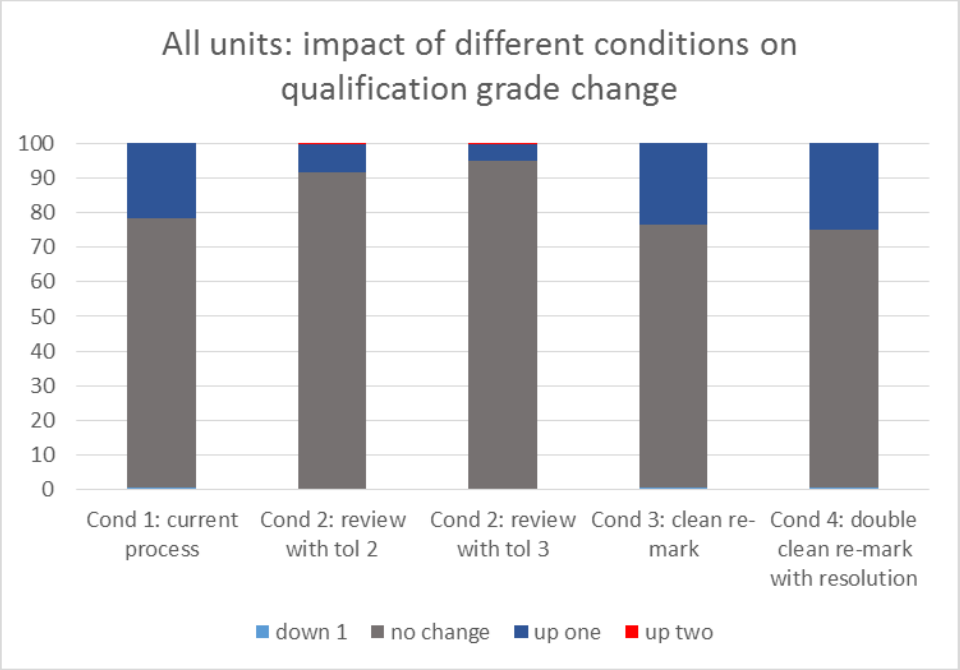


**Figure 5. Qualification-level grade change modelled according to different conditions for all units in the study**

## Proximity to true score for different models of EAR

The purpose of an EAR is to remedy marking error. Therefore, an important way to evaluate any possible method is in terms of the accuracy of the results provided. One way is to derive an estimate of the true score for each script and compare the extent to which different methods help markers generate marks on or close to the true score.

In the research, it was possible to estimate the true score for each script. This is the notional score that a candidate would get if there were no error in the measurement (as described above). In this study, we can estimate the true score by taking the average of all the marks for each script in condition 3 (single clean re-mark). Here, each script was marked by multiple markers, all independent of one another. We can then compare the proximity of any mark in any condition to the true score. The closer to the true score, the better – because it means the greater the accuracy of the mark.

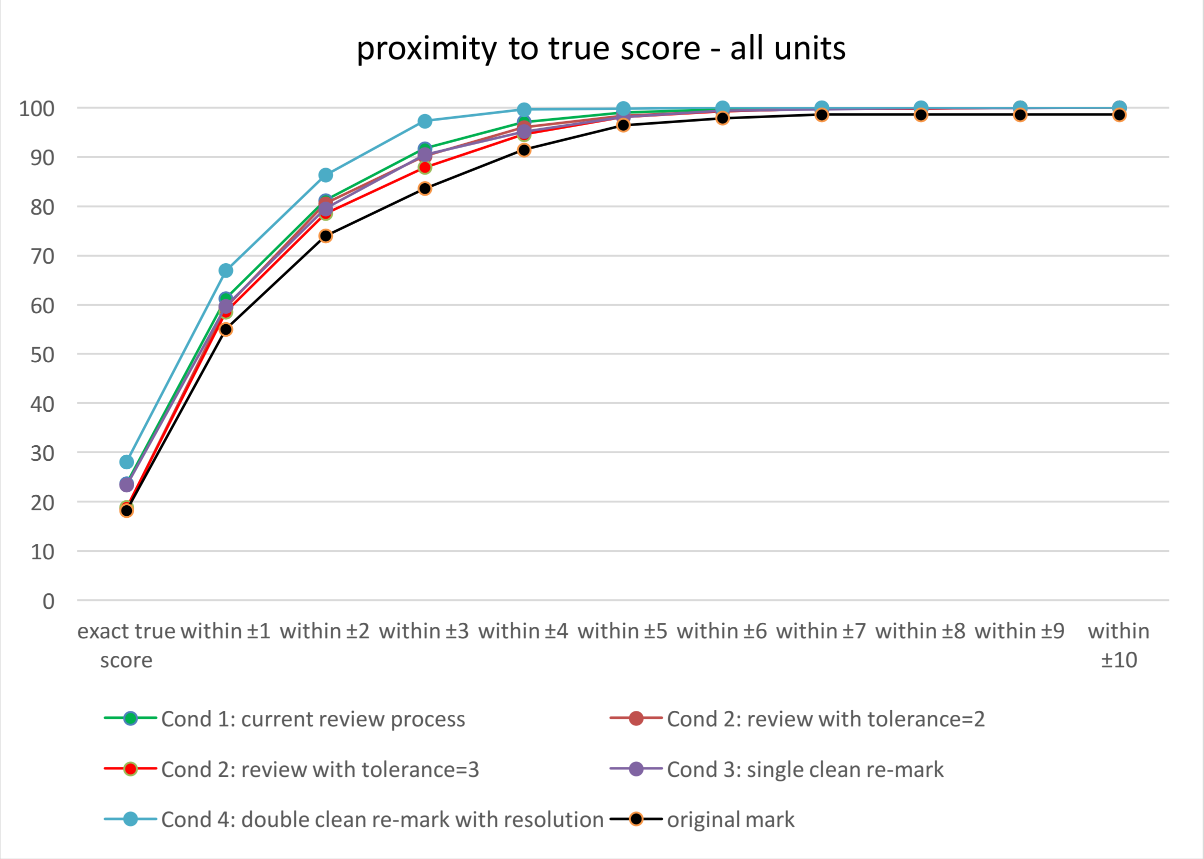


**Figure 6. Proximity to true score – a cumulative percentage chart for all units**

Figure 6 tells us what proportion of marks in each condition was exactly on or within a certain mark proximity. So, along the *x*-axis, we have the distance from the true score, starting with ‘exact’ (exactly on the true score), moving along to within ±1 mark, all the way to ±10 marks. The more marks that are exact or within a small margin of the true score, the better. The *y*-axis shows the percentage of scripts. Each condition, as well as the original mark, is indicated by a different colour line.

The key points from figure 6 are summarised as follows:

- We can see that 18 per cent of all the original marks (black line) for the sample of scripts in this study were exactly on the true score as derived in this study. This rises dramatically to 74 per cent within ±2 marks of the true score.

- All of the research conditions are associated with higher rates of proximity to the true score. This indicates that the research conditions and the different methods of reviewing and re-marking are successful in removing error to some degree.

- The condition that consistently has the highest rates of proximity to the true score is condition 4 (double clean re-mark with resolution) (blue line) – with 28 per cent of marks exactly on the true score and 86 per cent within ±2 marks of the true score.

- Condition 1 (current review process) (green line) gave results that were next best overall in terms of proximity to true score.

- Condition 2 (review plus tolerance) (orange and red lines) shows very interesting results. A tolerance of 3 marks gave lower rates of proximity to true score than a tolerance of 2 marks. This indicates that by using tolerance, we are getting further away from the true score and therefore adding error.

- Condition 3 (single clean re-mark) (yellow line) has overall produced outcomes quite similar to review plus tolerance in terms of proximity to true score.

There were some differences in the rank order of lines for individual units (see Appendix D).

## Summary of what the research tells us

We can summarise our findings as follows:

- The mark changes in the study were generally small (on average, less than 1 mark). The majority of mark changes (more than 85 per cent) in all units in all conditions were within 3 marks.

- The condition that is least likely to result in a mark or grade change is review plus tolerance (condition 2). However, the analysis with true score indicates that imposing a tolerance introduces error – probably because it suppresses a marker's expert judgement. In other words, tolerance would not be a fair method of preventing the substituting of one legitimate mark for another.

- The condition that produces outcomes closest to true score is double clean re-mark with resolution (condition 4). The margin between this condition and the next best – current review process (condition 1) – is around 5 to 6 per cent of scripts for up to within 4 marks' proximity to true score. This margin is noticeable but not substantial given the potential costs and difficulties of introducing such a model.

- Single clean re-mark (condition 3) involved too much variability and did not perform as well as some other conditions in terms of proximity to true score. It appears that the apparent advantage of markers being unbiased by the removal of the original marking might be more than offset by the risk of introducing some small errors.

# 5. Questionnaire findings

The questionnaire analysis is based on 31 respondents.

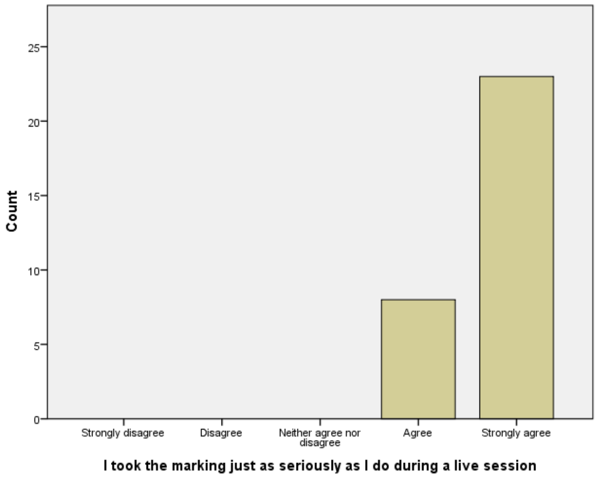
We asked about attitudes, views and behaviours in six distinct areas:

- A: Marking in the research study compared to marking in live sessions
- B: Live EAR marking
- C: Review plus tolerance
- D: Single clean re-mark
- E: Double clean re-mark with resolution
- F: Summing up of views on different processes.

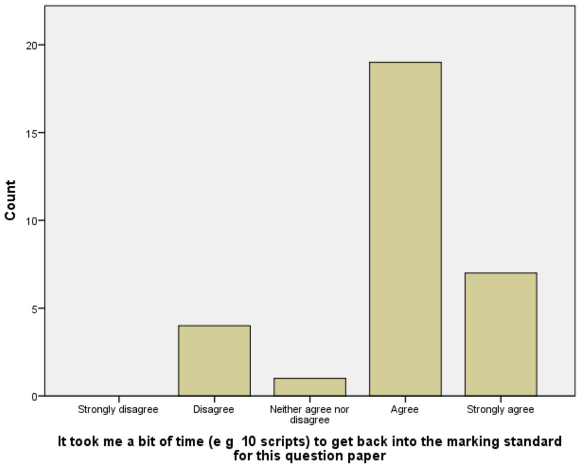
## A: Marking in the research study compared to marking in live sessions

The first set of questions were mainly about gaining some insight into the degree to which the examiners in the study took the marking seriously and marked or reviewed scripts authentically as they would do in a live, non-research context. Given the lapse of time since marking in the live session, paper-based rather than on-screen marking, as well as the knowledge of the work being research activity rather than live, it was important to gather some data from the examiners about how authentic they viewed their marking.

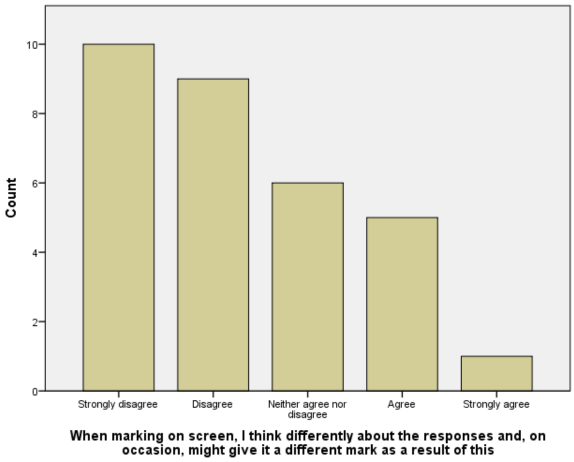
In general, the graphs in figure 7 provide a view that markers engaged with the marking tasks and marked authentically – 29 out of 31 respondents either strongly agreed or agreed with the following statement: ‘Overall the quality of my marking was very similar to that of a live session.’ Responses to other questions generally support a view that the marking was taken seriously and conducted as authentically as a live session.



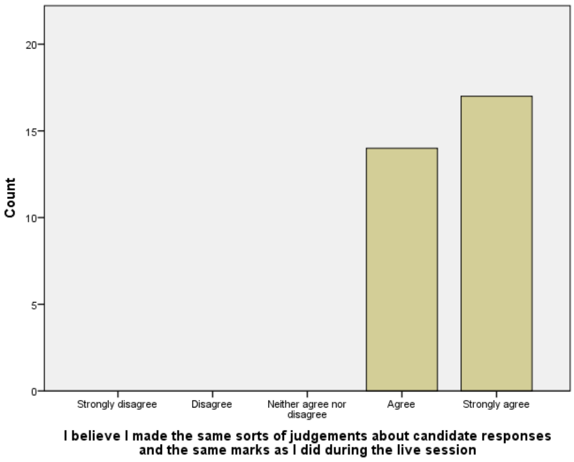
I took the marking just as seriously as I do during a live session.



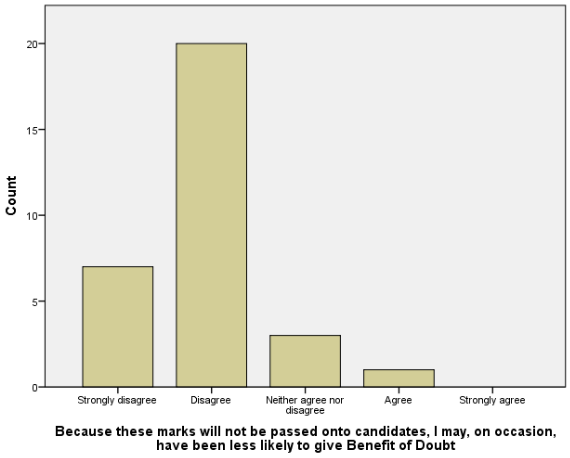
It took me a bit of time (e.g. 10 scripts) to get back into the marking standard for this question paper.

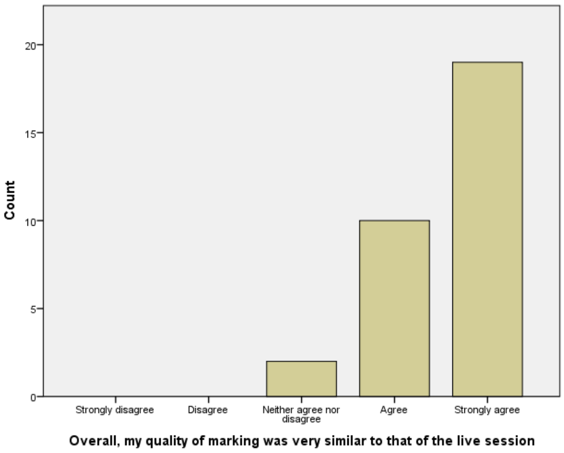


When marking on screen, I think differently about the responses and, on occasion, might give it a different mark as a result of this.



I believe I made the same sorts of judgements about candidate responses and the same marks as I did during the live session.





Because these marks will not be passed onto candidates, I may, on occasion, have been less likely to give Benefit of Doubt.

Overall, my quality of marking was very similar to that of the live session.

**Figure 7. Responses to a series of questions about the authenticity of examiners' marking in the research study**

Comments supporting this section were interesting. For example, one marker felt that marking on paper promoted better quality marking, while another felt that the lack of clearness of some photocopies may at times have reduced marking quality. Some markers used their original, personally annotated mark scheme; others did not or could not. Also relevant, although not mentioned here but mentioned in conversation with the researchers, several examiners were able to use their original standardisation materials to help remind them of the standard, while others no longer had these items. In one unit in particular, it was apparent that the standardisation scripts/items play an ongoing role in marking beyond that of the initial training and approval.

Comments that marking was very similar or the same:

| |
|---|
| I believe my marking is accurate, whether it is live or not. |
| The Mark Scheme is the ‘Bible’ and it is this that must be applied |
| Using my original mark scheme, rather than the clean copy provided, helped as it contained lots of additional info about marking decisions that I might otherwise have forgotten |
| I undertook this task as professionally as the live marking tasks. |
| Tried to replicate same conditions, treat marking in the same way as presumably results of this research could impact future EAR processes, so whilst not critical to candidates who sat these papers, could be critical to future candidates. |
| I prefer marking on paper, because it is physically less tiring. I marked according to the mark scheme, as always, and I re-marked according to our current rules – change a mark if you think it is wrong. |

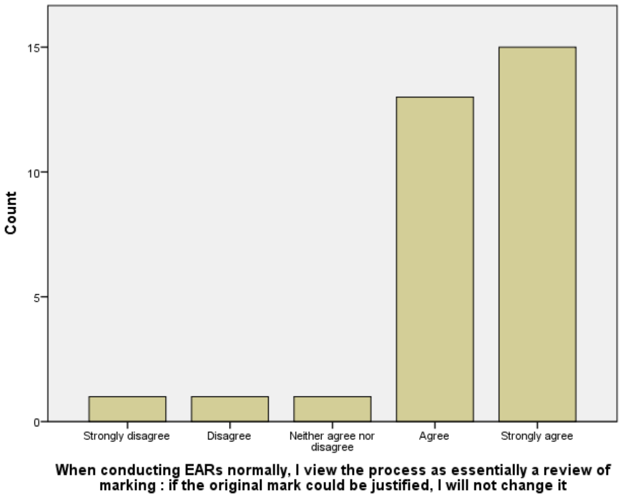
Comments about differences in the marking:

| |
|---|
| When marking on paper rather than on screen I find it much easier to make more of an accurate judgement with regards to the Extended Questions. This is not so much the case with the shorter questions as it is easier to apply the mark scheme. With the longer questions I am able to physically mark the points made by the candidates more easily. |
| As I no longer had my original mark scheme with added notes there were occasions when I may have awarded marks slightly differently. |

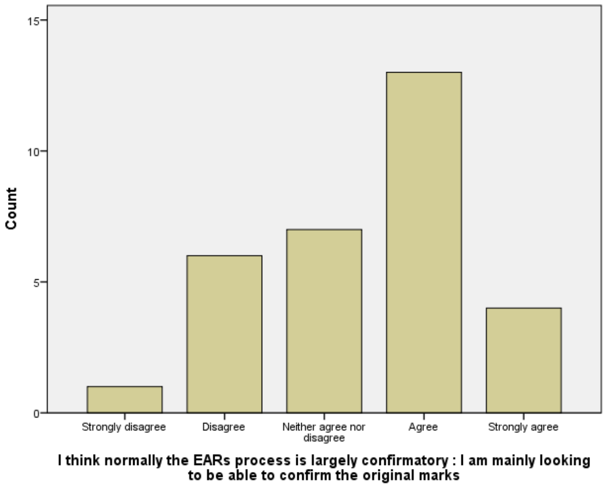
| |
|---|
| There were some scripts where the quality of the handwriting and photocopying made it difficult to decipher. On screen I would have been able to magnify, adjust the contrast which might have made the script more legible |
| Judgements made were different as the discussions around the mark scheme were no longer fresh in my mind. As a result my marking did differ to the live session. |
| Some judgements may have been affected by poor quality of the photocopies |
| I frequently discuss marks/queries with other Team Leaders. |
| When marking on line I mark by question but on paper it is more difficult to do that. I find marking by question improves the consistency of my decision making. |

## B: Live EAR marking

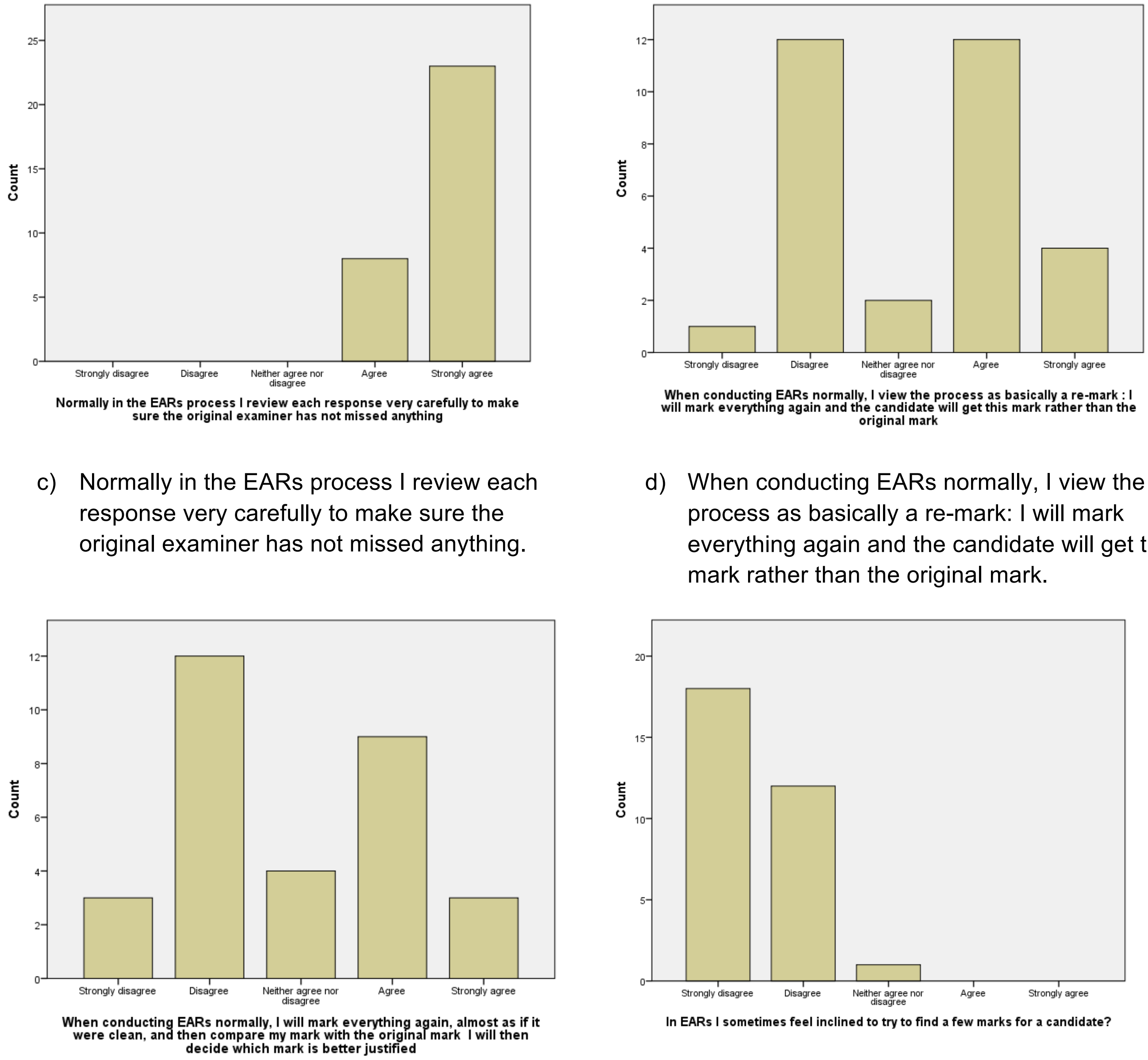
This set of questions looked at what examiners do in the normal (current) EAR process. The questions asked about the live EAR process, a review not a re-mark process – which all boards' instructions make clear to some extent. But the questionnaire aimed to uncover how examiners interpret this and how this translates into marking judgements. Overall, it is clear that there are some differences between examiners in how they interpret the 'review' instructions. In some cases examiners' reported marking behaviours appear to be divergent from the overall instructions to review the marking of the original examiner.



a) When conducting EARs normally, I view the process as essentially a review of marking: if the original mark could be justified, I will not change it.



b) I think normally the EARs process is largely confirmatory: I am mainly looking to be able to confirm the original marks.



c) Normally in the EARs process I review each response very carefully to make sure the original examiner has not missed anything.

d) When conducting EARs normally, I view the process as basically a re-mark: I will mark everything again and the candidate will get this mark rather than the original mark.

e) When conducting EARs normally, I will mark everything again, almost as if it were clean, and then compare my mark with the original mark. I will then decide which mark is better justified.

f) In EARs I sometimes feel inclined to find a few marks for a candidate.

**Figure 8. Examiner responses: attitudes and behaviours in live EAR (review) marking process**

There was much consensus around the following:

- Question (a) – the process as a review of marking
- Question (c) – reviewing each item response carefully
- Question (f) – **not** ‘finding a few marks for candidates’.

There was less consensus around other principles of marking EAR scripts.

The first question (a) asks about the extent to which examiners agree with the idea that the EAR process is a review – if the original mark can be justified, retain the original mark. Twenty-eight out of 31 agreed with this statement.

The second question (b) takes the idea of ‘review’ a bit further – to the idea that an EAR is ‘largely confirmatory’ – ‘I am mainly looking to be able to confirm the original marks’. Seventeen examiners agreed or strongly agreed with this statement, seven disagreed or strongly disagreed, and seven were undecided. The variability of responses on this item might suggest slightly different reviewing/marking practices between examiners.

The particularly interesting item is around re-mark rather than review (question d). A re-mark implies that, although the original mark may be visible, it is, for at least part of the EAR process, ignored. The examiners were polarised on this, with a bimodal distribution – 12 examiners agreeing and 12 examiners disagreeing. Without further questioning, it is not clear how a re-mark strategy is compatible with a ‘review’ strategy (question (a)), especially as question (d) states: ‘I will mark everything again and the candidate will get this mark rather than the original mark’. Possibly, examiners do not perceive a difference in the language and or concomitant behaviours between ‘review’ and ‘re-mark’. Indeed, in two out of three boards’ EAR examiner instruction documents (from 2014), these words were used interchangeably on one or more occasions.

The next item on the questionnaire (question (e)) represents a hybrid between re-mark and review: ‘I view the process as basically a re-mark – I will mark everything again, almost as if it were clean, and then compare my mark with the original mark. I will then decide which mark is better justified.’ There is again some polarisation in responses here, but with slightly fewer respondents agreeing or strongly agreeing compared to question (d) (12 respondents compared to 16). It is also worth pointing out that there was no association between examiners’ responses to questions (d) and (e). In other words, a ‘re-marking’ examiner in question (d) might also adopt the hybrid approach described in question (e) or they might not.

A picture emerges, then, of different examiners applying different principles around marking EAR scripts/items. There is not a consistent reported approach across all examiners, even within a board or unit. Only 31 examiners answered the questionnaire and it is not known how this generalises to the whole population of EAR markers (likely to be several thousand across all boards and all examined units). However, the fact that these findings are not unit or board specific suggests that it could be a wider issue for more examiners. This suggests, then, that there might be an element of examiner lottery at EAR such that if a script is allocated to an examiner who takes a ‘reviewing’, a ‘re-marking’ , or a ‘hybrid’ approach, the outcome may have a differing probability of resulting in a changed mark.

We also asked examiners about eight different marking scenarios, in order to understand how particular principles might be related to more specific judgements around changing versus retaining marks.

The scenarios for points-based mark schemes indicated whether or not the EAR examiner would have in mind a higher or lower mark than the original, as well as if the original mark ‘could’ or ‘could not be justified’.

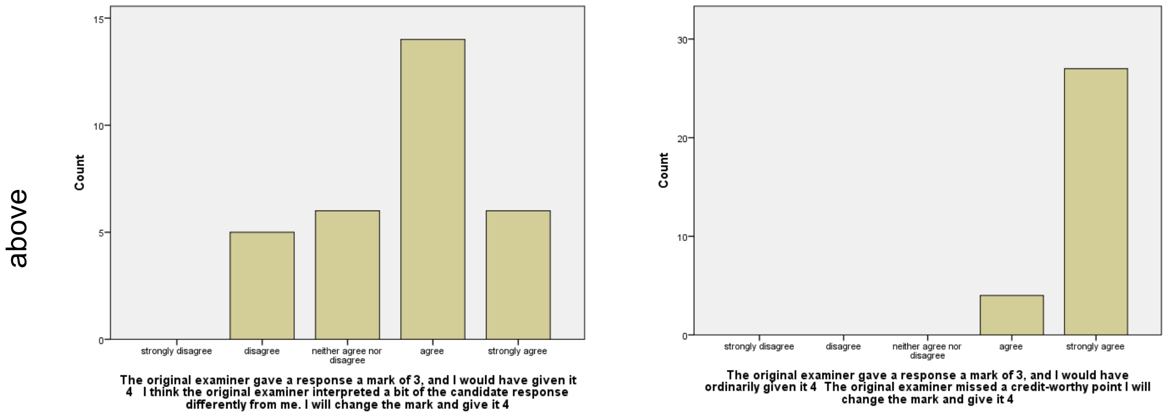
In each scenario, examiners rated statements according to a five-point scale:

- Strongly agree – I would not have done this.
- Disagree – I probably would not have done this.
- Neither agree nor disagree – I don’t know.
- Agree – I probably would have done this.
- Strongly agree – I definitely would have done this.

From figure 9, we can see that only one question produced strong consensus – that around raising a mark where the examiner had missed a creditworthy point. The other scenarios had less consensus but there is some sort of suggestion that even where judgement and interpretation are involved, examiners would prefer to change the mark upwards.

Original mark can be justified (judgement/interpretation)

Original mark difficult to justify ('error')

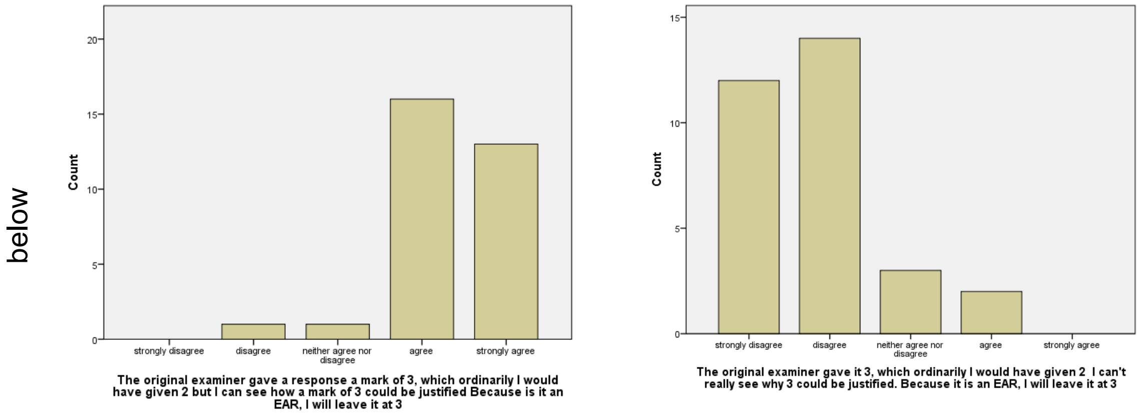


The original examiner gave a response a mark of 3, and I would have ordinarily given it 4. I think the original examiner interpreted a bit of the candidate response differently from me. I will change the mark and give it 4.

'Correct' answer: strongly disagree/disagree

The original examiner gave a response a mark of 3, and I would have ordinarily given it 4. The original examiner missed a creditworthy point. I will change the mark and give it 4.

'Correct' answer: strongly agree



The original examiner gave a response a mark of 3, which ordinarily I would have given 2 but I can see how a mark of 3 could be justified. Because it is an EAR, I will leave it at 3.

'Correct' answer: strongly agree/agree

The original examiner gave it 3, which ordinarily I would have given 2. I can't really see why 3 could be justified. Because it is an EAR, I will leave it at 3.

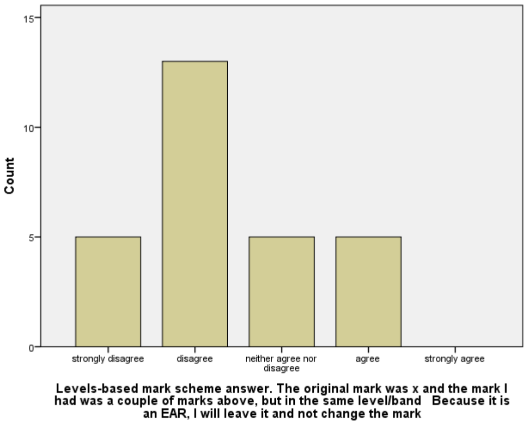
'Correct' answer: strongly disagree/disagree

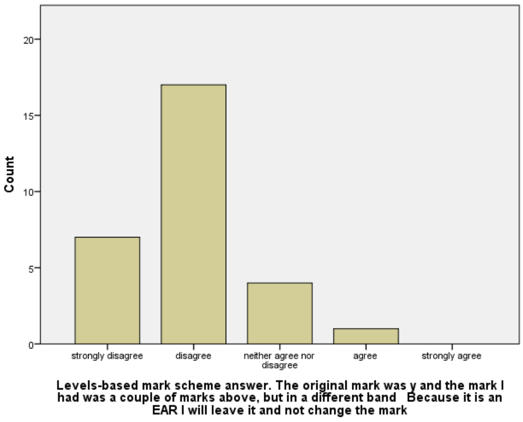
**Figure 9. Points-based scenarios and whether or not examiners would change marks according to whether they were above or below original and whether differences are not justifiable ('error') or could be justified ('judgement / interpretation')**

Within band

Outside of band

Above





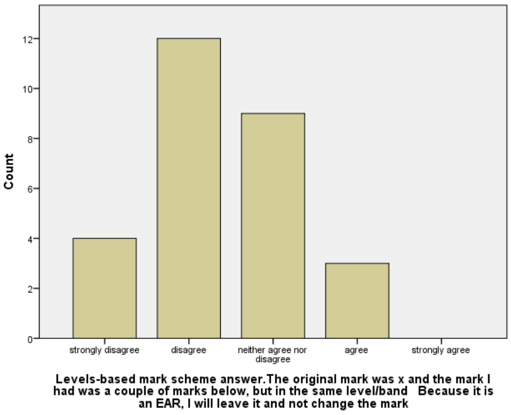
The original mark was x and the mark I had was a couple of marks above, but in the same level/band. Because it is an EAR, I will leave it and not change the mark.

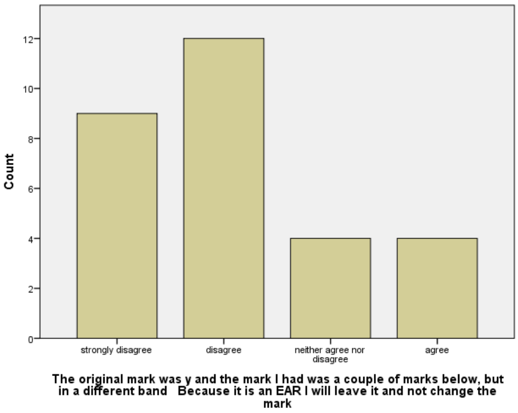
‘Correct’ answer: strongly agree/agree

The original mark was y and the mark I had was a couple of marks above, but in a different band. Because it is an EAR, I will leave it and not change the mark.

‘Correct’ answer: probably disagree (less clear, depending on scenario)

below





The original mark was x and the mark I had was a couple of marks below, but in the same level/band. Because it is an EAR, I will leave it and not change the mark.

‘Correct’ answer: strongly agree/agree

The original mark was y and the mark I had was a couple of marks below, but in a different band. Because it is an EAR, I will leave it and not change the mark.

‘Correct’ answer: probably disagree (less clear, depending on the scenario)

**Figure 10. Levels-based scenarios and whether or not examiners would change marks according to whether they were above or below original and within or not within same band/level**

The responses for scenarios around levels-based mark schemes are difficult to interpret. Generally speaking, levels-based mark schemes are likely to deal with more subjectively marked responses and therefore it can be difficult to justify small mark differences within band as requiring mark changes as both marks are likely to be ‘legitimate’. It does, however, indicate that examiners, despite being in the territory of what is likely to be judgement, are, in all scenarios, more inclined to change the mark rather than leave the original mark – and they are slightly more inclined to change marks where this would involve an upward change rather than downward change.

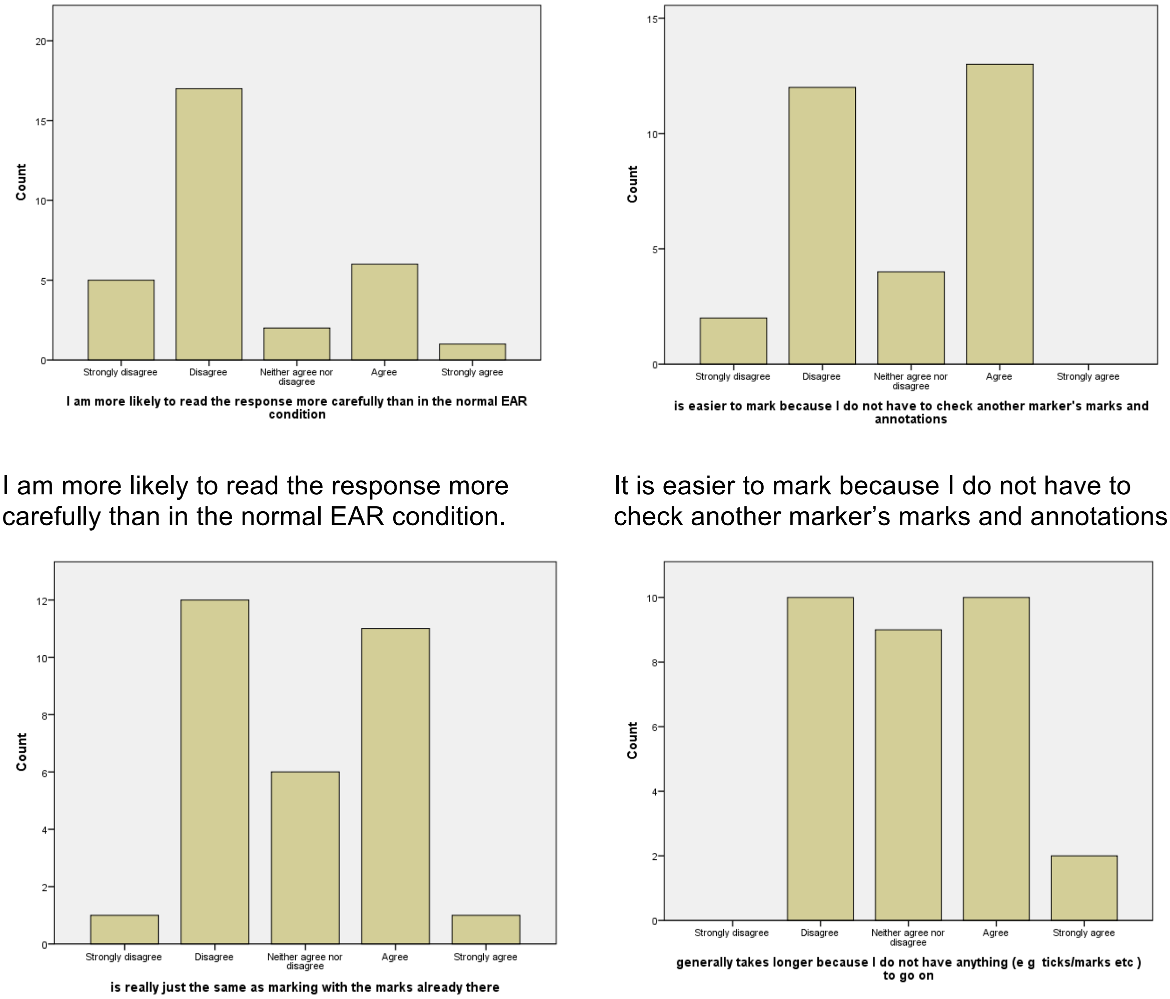
## C: Review marking with tolerance

Markers were not told the value of tolerance that would be applied. This meant that the researchers could apply different values of tolerance at script level. However, it might have resulted in markers believing or assuming different things. Respondents were asked whether or not knowing that tolerances would be applied had any impact on how they awarded marks compared to the standard EAR condition. Twenty-eight of the 31 respondents answered no. The three respondents who answered yes explained their responses as follows:

| |
|---|
| Where purely ‘opinion’ more likely to go with my mark. |
| Where a candidate’s responses had been marked harshly by the original examiner I felt the candidate deserved more credit. However, if the tolerance rule was applied the candidate is unlikely to have the mark increased which I felt unjust. I therefore was more likely to alter borderline decisions in this style of EAR. |
| in questions with tolerances of 2, I would be less likely to change a mark if the original was within 1 mark of my own |

## D: Single clean re-mark

A series of four questions asked about blind or clean re-marking in comparison with marking scripts with marks (and annotations) visible. These questions tap into attentional and/or attitudinal aspects of marking EAR scripts with no visibility of the original marks. The responses are shown in figure 11.



I am more likely to read the response more carefully than in the normal EAR condition.

It is easier to mark because I do not have to check another marker's marks and annotations.

It is really just the same as marking with the marks already there.

It generally takes longer because I do not have anything (e.g. ticks/marks) to go on.

**Figure 11. Responses to questions about single clean re-mark**

Again, there is little consensus between examiners – and this was not generally because of board or unit differences. It therefore seems likely that these differences stem from individual differences in examiners' marking/reviewing processes – for example, how much attention they pay to the original marks and annotations and/or to the candidate response.

## E: Double clean re-mark with resolution

This section asked about the process of double marking and the resolution process.

All examiners were asked to resolve all scripts with any mark discrepancies (at item level as well as at whole script level). Each examiner was paired with another examiner for one or two packs of scripts, each pack containing the equivalent of ten scripts). After seeing a file identifying on which items/scripts they had discrepancies, each pair arranged a resolution phone call. Responses from the questionnaire indicated that this took between 30 minutes to over one hour per pack, with 11 examiners indicating over one hour per pack.

Examiners were also asked about the reasons why two examiners might have different marks for any particular item (see figure 12).

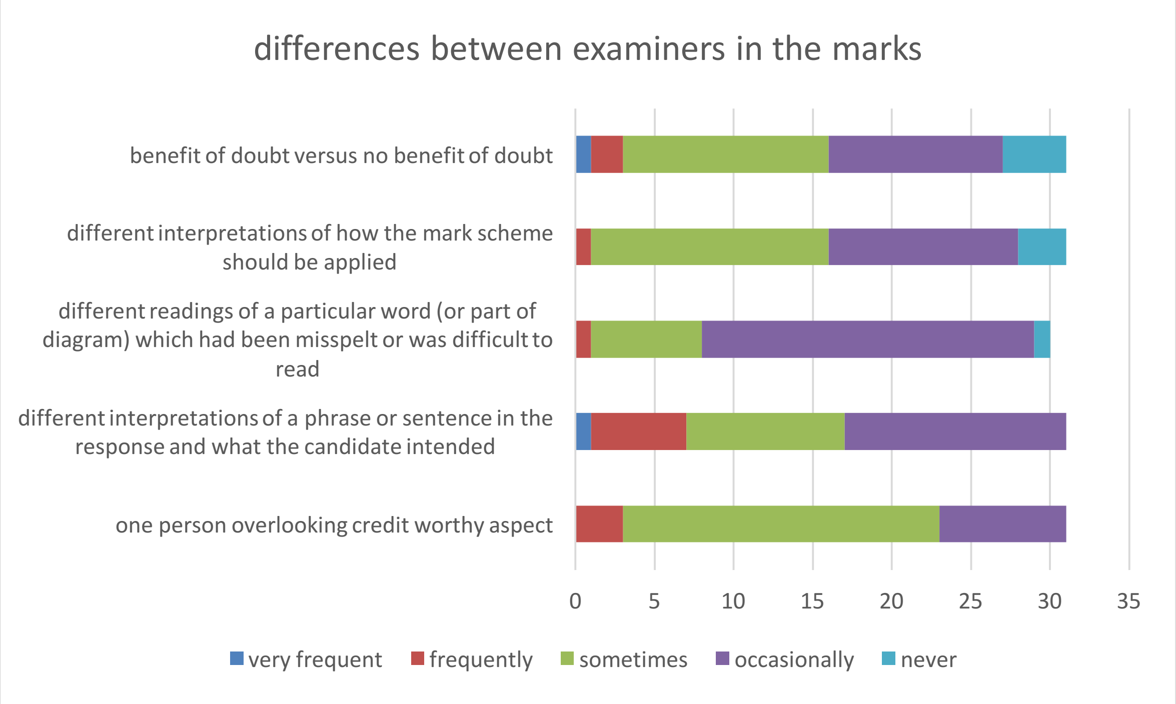


**Figure 12. Reasons for differences between examiners' marks in double clean re-mark with resolution**

The reason given for markers' discrepancies with the highest combination of 'very frequent' and 'frequent' was 'different interpretations of a phrase or sentence in terms of what the candidate probably intended'. This is a reminder that candidates do not necessarily write using the language of the mark scheme and sometimes may not express themselves with perfect clarity. Even in subjects with points-based mark schemes, this can leave room for differences of marker opinion of 1 mark or more.

The reason with the highest combination of 'very frequent', 'frequent' and 'sometimes' was 'one person overlooking a creditworthy aspect of an answer'. This might indicate an omission on one examiner's behalf, or possibly an issue more of interpretation as to what constitutes creditworthy material.

Examiners were also asked about any other reasons for items with discrepant marks, with the following responses:

- Interpretation of the mark scheme (n = 3)
- Lack of clarity in the mark scheme (n = 1)
- Misapplication of the marking criteria (n = 2)
- Not crediting a correct response not mentioned on the mark scheme (n = 1)
- Some markers having their original versions of the mark schemes with additional notes that allowed or disallowed some responses (n = 1)
- An individual or unusual response from the candidate (n = 2)
- Inability to read the candidates' handwriting (n = 1)
- An oversight (n = 1).

We were also interested in how the resolution process worked – the extent to which it was straightforward or difficult and what factors came into play. Responses are summarised as follows:

- Examiners reported it was 'very easy' (n = 6), 'quite easy' (n = 14) or 'middling' (n = 11) to come to an agreed view for each item. No examiners reported it as being generally 'difficult' or 'very difficult'.
- Twenty-eight examiners provided additional comments about the sorts of things that were more difficult to resolve. Nearly all the responses included reference to either different interpretations of candidate responses and/or how these might fit with different interpretations of the mark scheme – usually on very specific and 'technical' points, and usually around very small mark differences (usually just 1 mark).
- Three examiners reported that particular responses were impossible or very difficult to resolve. They said that the difficulty to come to a joint view on the most appropriate mark lay in differing interpretations of the mark scheme and/or candidate response.

- The overall script mark was generally not taken into account in resolving marks – only two examiners reported referring to or looking at the overall script mark. The implication is that markers made decisions on an item-by-item, response-by-response basis.

- Trading off 'benefit of doubt' and 'no benefit of doubt'[14] within a script was noted by seven examiners who marked EAR whole scripts in this study.

Examiners supplied interesting comments around how benefit of doubt (BOD) and no benefit of doubt (NBOD) worked in the resolution process. Some of these are included below – not least because they provide useful insights into the greyer areas of making judgements around single marking points in marking scripts.

| |
|---|
| If one examiner had given a BOD for one question, very occasionally it was agreed to give another BOD, particularly where the candidate did not always necessarily use the correct terminology, but had shown some understanding of the question, which was down to interpretation by the marker. |
| Due to difference of opinion we felt it fair that we gave a BOD and NBOD. |
| During discussion with another examiner this was used to reach a compromise. |

## F: Summing up of views on different processes

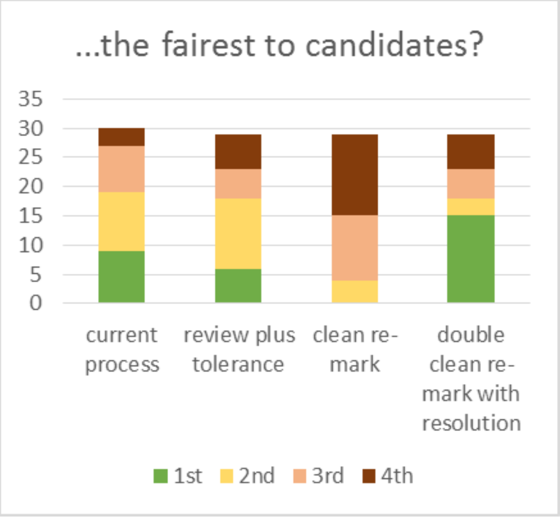
In the last section, examiners were asked about their own views on the different conditions. This allowed examiners to share some of their own insights from having conducted all four marking conditions.

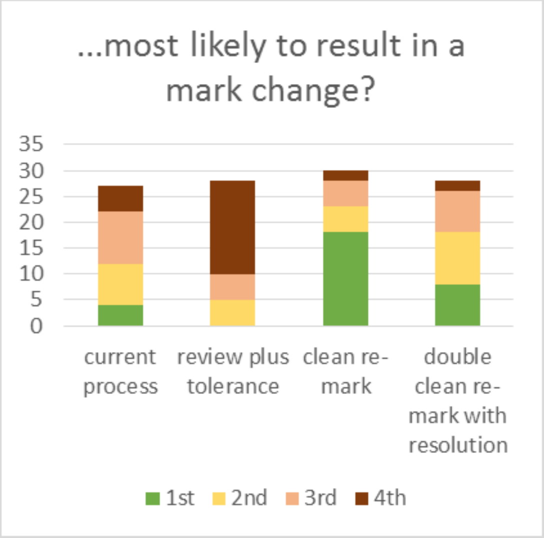
Many of the views summarised in the figures (see figure 13) reflect the findings from the experimental part of this research study. Double clean re-mark with resolution was ranked top for most accurate and the fairest for candidates. In contrast, single clean re-mark ranked lowest overall against these two criteria. This seems to indicate that examiners themselves do not trust this process in comparison to the other processes, but that their confidence in fairness and accuracy is significantly

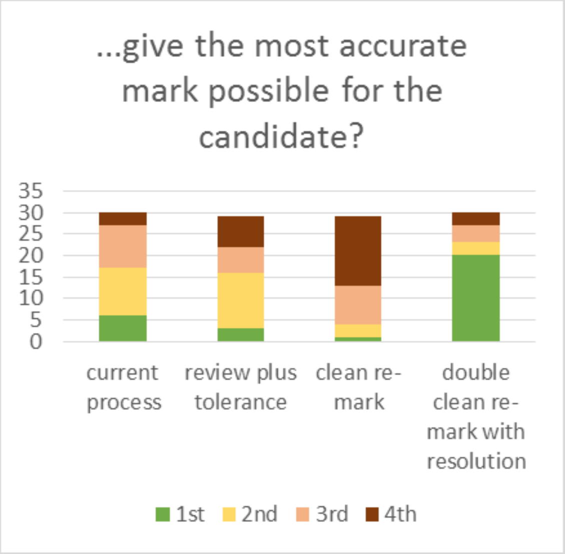
[14] Benefit of doubt is the concept that where a candidate response has not quite demonstrated beyond doubt the required knowledge or understanding, an examiner might feel that it is close enough to the creditworthy answer in the mark scheme. Sometimes this might be where a candidate has expressed themselves slightly unclearly or perhaps has apparently accidentally omitted a word in a sentence. Anecdotally, so that a candidate cannot 'over-benefit' from benefit of doubt, where an examiner is marking whole scripts they may adopt a compensatory approach such that a candidate who receives benefit of doubt on one occasion within a script will not receive it on the next occasion.

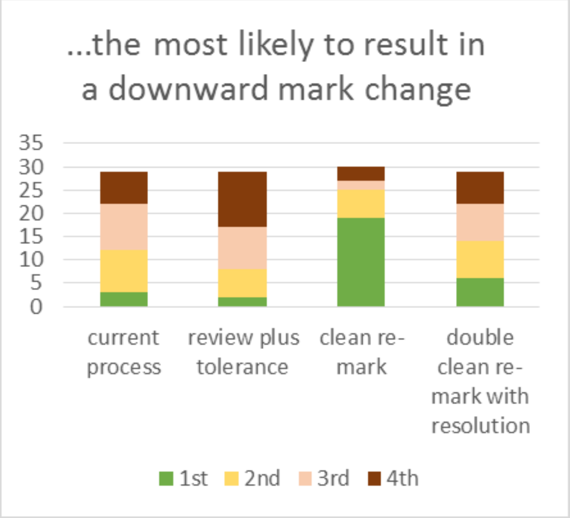
increased (transformed, even) through the resolution process with another examiner. This point is picked up again in later examiner comments.

*Please place the conditions in order according to which condition is, in your view, …*









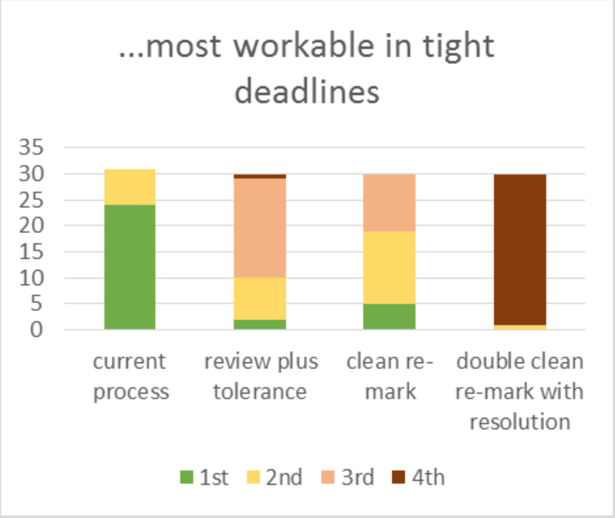


**Figure 13. Examiners’ views of the different processes – rank ordering against different criteria**

Single clean re-mark was ranked as most likely to result in a mark change and downward mark change. This was also reflected in the experimental data. Finally, the current process was regarded as most workable in tight deadlines, with double clean re-mark with resolution judged as the least workable. Examiners would have been aware of the time taken to discuss scripts (approximately six minutes per script), as well as additional administrative time around setting up times to discuss with examiners and ensure that a final marked version of script is completed.

The final closed item on the questionnaire asked examiners to select the process that they thought was best overall for the future (see figure 14).

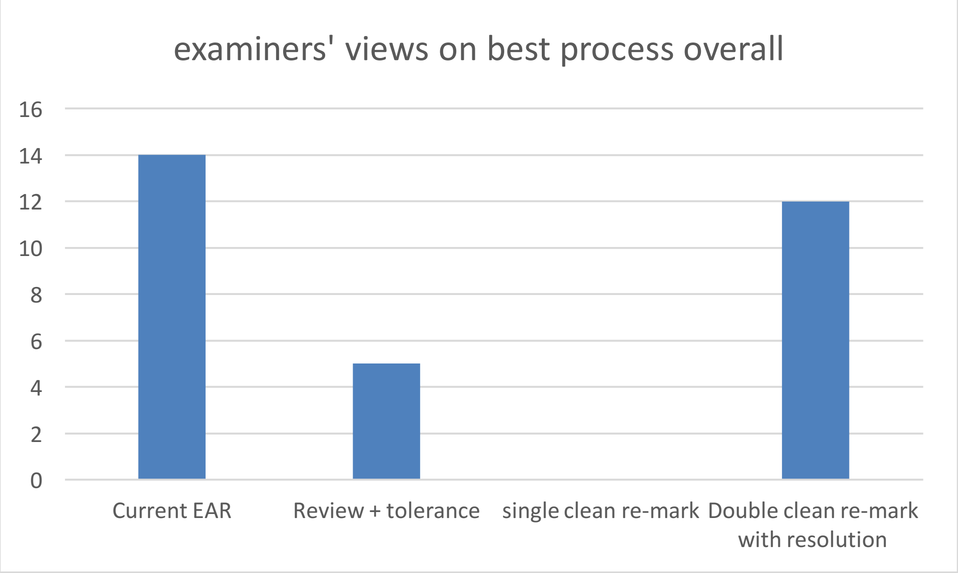


**Figure 14. Examiners’ views on the best process overall for the future**

Figure 14 shows that the current EAR process was selected by 14 of the 31 examiners, closely followed by double clean re-mark with resolution. No examiner selected single clean re-mark.

Examiners supported their choice with reasons. Some of these are given below as the insights are valuable given their experience with all four conditions.

Reasons for preferring current EAR:

| |
|---|
| Being able to see the original mark gives me confidence that I haven't missed anything that the original marker saw. |
| Close review of original marking taking annotation into consideration and close adherence to marking criteria should affirm or otherwise the mark awarded. The |

| |
|---|
| more interpretations considered will make EAR procedures unmanageable and the time needed impractical. |
| Double blind with resolution sounds good but resolution will be impacted on by personalities of those ‘discussing’ how marks should be resolved – a more forceful personality having potentially undue influence. |
| Fairest approach and has marks already visible so can see where marks have been awarded by another examiner. Also speeds up the marking process. |
| Fastest and most economical. If you add any more blind marks to [this sort of] paper you will just get more problems, as there is no one correct answer. |
| Given the tight deadlines that surround EAR scripts the current process allows the reviewing examiner to see the original marks and make their own interpretation. When conducting blind remarks it is possible to make mistakes – these are less likely with the current EAR process. Condition 4 would yield the most reliable mark but is impossible in the deadlines we have to work to. |
| I don't think any of the other options are solid enough, in terms of how they would be operationalised, and on that basis, at least the current system is operationally as nuanced as it could be because of the organisational experience contained within it, which maximises the number of checks and balances likely to be in play…. |
| It seems to be mainly effective and is not as time consuming as the double blind remark. |
| This works well for [my subject] although for subjects that are more subjective like English/essay based subjects I think condition 4 would be better. |

Reasons for supporting review plus tolerance model:

| |
|---|
| As it is a mix of old and new. If there are major differences a blind remark can occur to allow an accurate mark to be produced. |
| First, I do not think there is one best way. A range of services should be offered from the current EAR to a re-mark something like C4. The more radical remarks would be more costly and time-consuming… |
| This process still allows for slight natural professional differences. |

Reasons for supporting double clean re-mark with resolution often pointed to fairness and fail-safeness, but frequently pointed out practical or logistical difficulties that might occur if implemented in a live session. Reasons included:

| |
|---|
| a fairer way to reach an undisputed result |
| It highlighted accidental errors in the marking and therefore gave the most accurate mark. I'm not sure how feasible the resolutions would be though on a large scale. |
| Fairest and most accurate result. |
| Fairest for the candidate as it does not rely on just one interpretation of their answers. |
| I think this was a very fail safe method, but I think it is also unworkable in terms of time and cost. For example, to resolve a pack of 10 scripts over the phone took two experienced colleagues an hour, plus it took me about almost 3 hours to mark the packs in the first place… I don't think the exam board would pay a rate that reflects that time given we often have high volume in a short space of time. And they would definitely not pay for the phone calls – we have trouble claiming for telephone calls now!! |
| Often candidates answer in a manner that does not easily fit the mark scheme. This gives the opportunity to discuss with another experienced professional the fairest judgement to be applied. |
| Paper will have been reviewed more times and by discussing marks differences a more informed decision should result. |
| This is a very fair way of doing things but it is incredibly time consuming and would be difficult when meeting strict deadlines. |
| Two experienced examiners marking separately and then discussing and arriving at an agreed mark without being influenced by the original mark seems the most fair. How practical this is given the quantity of remarks, the narrow time-frame and cost remains a moot question though. |

# 6. Discussion

The research has been valuable in exploring the impact of different potential EAR processes on mark and grade changes as well as the accuracy of results. While the review plus tolerance model had had perceived advantages before the research in terms of apparently preventing the replacement of one legitimate mark with another, it appeared that this was at the expense of accuracy since applying a tolerance reduced the proximity to the true score. This suggests that using tolerance introduces an element of error into results by negating or ignoring the expert judgement captured in the review process. It follows, then, that any process for preventing one legitimate mark being replaced by another has to rely on expert judgement rather than a fixed, unseeing, numeric rule.

Double clean re-mark with resolution had better proximity to true score overall, but only by a few percentage points. Gains in accuracy of this magnitude are consistent with other research on double marking (e.g. Fearnley, 2005). Work conducted and reported by us elsewhere[15] informed a consideration of the extent to which any advantage gained in this model might be undone by costs and impact of implementation. The examiners in the study, even those who preferred this model overall, pointed out some significant barriers to implementation in a live EAR session.

Single clean re-mark was associated with the greatest rates of and variability in mark change. Overall, this method gave proximity to true score just behind that of the current review process. There were some unit differences in terms of relative proximity to true score for this condition. While for units A and B, single clean re-mark performed better than review plus tolerance, it performed worst of all for unit C. This may be due to the nature of the assessment in this unit, with a high proportion of high-tariff items. Such items are particularly prone to marker variability due to the scope for interpretation of candidates' responses.

One potential limitation of such research is that it takes place outside of the context of a live marking or EAR process and some time has passed since the main marking period. There is some potential for the marking standard to alter or for examiners to take the marking less seriously than in a live session. An additional issue in this research is that it was conducted using paper marking, whereas the majority of marking for these units will be conducted online. In order to mitigate the impact of these risks, examiners were required to 're-standardise' before they could progress on to the main part of the study. While it is difficult to be certain that the research context overall did not compromise the authenticity of the marking, there are two indicators that there was only a minimal effect. First, examiner responses to the

[15] Regulatory Impact Assessment - https://www.gov.uk/government/consultations/marking-reviews-appeals-grade-boundaries-and-code-of-practice

questionnaire generally supported the notion of authentic engagement with the task. Second, the research outcomes for the current process were very similar to the actual live EAR process for the same scripts, but with a difference in only one respect – that markers in research were less likely to retain the overall script mark and slightly more likely to remove marks. This was true for all three units in the study. We can only speculate why this difference is present in the research. There are two main hypotheses. One is that markers in research, some months later and marking on paper rather than on screen, became slightly more severe in the research generally (although this is not reflected in the upward mark figures). Alternatively, in the research scenario, markers are more inclined to remove marks where it is justified because they know this cannot impact upon the actual candidate. It is possible, then, in the live scenario, that markers may feel reluctant to remove marks because they are mindful of the potential negative impact on the candidate. This possible interpretation reminds us that markers undertaking EAR may be making their judgements in the context of being aware of the possible ‘plight’ of the candidate whose script they are reviewing or re-marking.[16] This may in some cases exert some influence on their decisions whether or not to remove a mark.

While this research has some limitations in terms of relatively small numbers of scripts and examiners, the consistency of findings for the different processes across three quite different units gives a certain strength to the findings overall. The insights from the research have very much informed the proposals in the consultation.

Data from the questionnaire indicates that for the current EAR process (a review), some examiners are adopting slightly different processes – some indicated that they were reviewing the marking of another examiner, while others said that they adopted some elements of re-mark process, ignoring the original marks. This suggests that there could be some gains in consistency of practices between examiners through better guidance, instructions and training for examiners.

[16] Although examiners at EAR would not know details of, say, the candidate’s name or overall qualification grade, they could likely infer that a candidate was not content with their result.

# 7. Summary of findings and conclusions

- The mark changes in the study were generally small (on average, less than 1 mark). The majority of mark changes (more than 85 per cent) in all units in all conditions were within 3 marks.

- The condition that is least likely to result in a mark or grade change is review plus tolerance (condition 2). However, the analysis with true score indicates that imposing a tolerance introduces error – probably because it suppresses a marker's expert judgement. In other words, tolerance would not be a fair method of preventing the substituting of one legitimate mark for another.

- The condition that produces outcomes closest to true score is double clean re-mark with resolution (condition 4). The margin between this condition and the next best – current review process (condition 1) – is around 5 to 6 per cent of scripts for up to within 4 marks' proximity to true score. This margin is noticeable but not substantial given the potential costs and difficulties of introducing such a model.

- The current review process (condition 1) performed well in terms of proximity to true score, second only to double clean re-mark with resolution.

- Single clean re-mark (condition 3) involved greater mark variability and did not perform as well as some other conditions in terms of proximity to true score. It appears that the apparent advantage of markers being unbiased by the removal of the original marking might be more than offset by the risk of introducing some (new) small errors. Examiners themselves also rated this method as least likely to deliver fair and accurate results.

- Data from the questionnaire indicates that for the current EAR process (review), some examiners are adopting slightly different processes. Some examiners indicated that they were reviewing the marking of another examiner, while others indicated that they adopted some elements of the re-mark process. Greater guidance and training is probably needed in order to ensure greater consistency of approach.

# 8. References

Billington, L. (2012) *Exploring Second Phase Samples: What is the Most Appropriate Basis for Examiner Adjustments?* Manchester: AQA, Centre for Education Research and Policy.

Black, B., Suto, 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 & Practice* 18(3), 295–318.

Fearnley, A. (2005) *An Investigation of Targeted Double Marking for GCSE and GCE*. London: QCA.

HMC (Headmasters and Headmistresses Conference) (2012) *England's 'Examination Industry': Deterioration and Decay. A Report from HMC on Endemic Problems with Marking, Awarding, Re-Marks and Appeals at GCSE and A Level 2007–12*. Available at: www.hmc.org.uk/wp-content/uploads/2012/09/HMC-Report-on-English-Exams-9-12-v-13.pdf

JCQ (2015) *GCSE, GCE, Principal Learning and Projects (including Extended Project): Post-Results Services. Guidance for Centres*. London, UK. Available at: http://www.jcq.org.uk/exams-office/post-results-services

Meadows, M.L. and Baird, J. (2005) *What is the Right Mark? Respecting other Examiners' Views in a Community of Practice*. Poster presented at the AEA Europe conference in Dublin, November 2005. Cited in Fearnley (2005).

Murphy, R.J.L. (1979) Removing the Marks from Examination Scripts before Re-Marking Them: Does It Make Any Difference? *British Journal of Educational Psychology* 49, 73–78.

Vidal Rodeiro, C (2007). *Agreement between outcomes from different double-marking models.* Research Matters: A Cambridge Assessment Publication, 4, 28-34.

# Appendix A: Script sampling for each unit

**Table A1: Unit A scripts in live EAR session compared to scripts in the study**

| | All EAR scripts in 2014 session (n > 3,000) | | | | EAR scripts in study (n = 120) | | | |
|---|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | SD | Min | Max | Mean | SD |
| Original unit raw mark | 3 | 68 | 42.5 | 11.0 | 15 | 68 | 42.30 | 11.0 |
| Original qualification UMS | 28 | 179 | 137.1 | 23.8 | 79 | 179 | 135.9 | 22.0 |
| EAR unit raw mark | 3 | 70 | 42.9 | 11.2 | 15 | 68 | 42.3 | 11.0 |
| Mark difference (EAR − original) | −13 | 10 | .34 | 1.7 | −13 | 10 | 0.39 | 2.3 |

**Table A2: Unit B scripts in live EAR session compared to scripts in the study**

| | All EAR scripts in 2014 session (n > 1,000) | | | | EAR scripts in study (n = 80) | | | |
|---|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | SD | Min | Max | Mean | SD |
| Original unit raw mark | 1 | 89 | 57.35 | 15.502 | 17 | 80 | 55.91 | 14.676 |
| Original qualification UMS | 52 | 479 | 225.08 | 86.539 | 85 | 471 | 214.74 | 72.890 |
| EAR unit raw mark | 1 | 89 | 57.65 | 15.567 | 17 | 82 | 56.85 | 14.493 |
| Mark difference (EAR − original) | −5 | 44 | 0.3 | 2.043 | −3 | 44 | 0.94 | 5.032 |

### Table A3: Unit C scripts in live EAR session compared to scripts in the study

| | All EAR scripts in 2014 session (n > 4,000) | | | | EAR scripts in study (n = 80) | | | |
|---|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | SD | Min | Max | Mean | SD |
| Original unit raw mark | 9.0 | 66.0 | 48.9 | 4.5 | 20.0 | 58.0 | 48.7 | 5.6 |
| Original qualification UMS | 56.0 | 209.0 | 173.8 | 10.0 | 116.0 | 209.0 | 173.2 | 11.9 |
| EAR unit raw mark | 7.0 | 67.0 | 49.8 | 4.6 | 25.0 | 60.0 | 49.6 | 5.5 |
| Mark difference (EAR − original) | −9 | 15 | 0.89 | 2.1 | −9 | 15 | 0.9 | 2.9 |

# Appendix B: Mark changes for each condition for units B and C

### Table B1: Summary of mark changes (unit B)

| Condition | | Largest negative mark change | Largest positive mark change | Mean mark change (later mark − original mark) | SD |
|---|---|---|---|---|---|
| Baseline | Actual EAR process | −3 | 44 | 0.93 | 5.01 |
| Condition 1 | Current review process | −8 | 44 | 0.66 | 5.12 |
| Condition 2 | Review plus tolerance (marks as submitted before tolerance applied) | −6 | 45 | 0.58 | 5.00 |
| Condition 3 | Single clean re-mark | −8 | 45 | 0.73 | 5.14 |
| Condition 4 | Double clean re-mark with resolution | −8 | 44 | 0.46 | 5.12 |

**Table B2: Percentages of mark changes (unit B)**

| Condition | | % scripts with negative mark change | | % no change | % positive change |
|---|---|---|---|---|---|
| Baseline | Actual EAR process | 17.50 | | 42.5 | 40 |
| Condition 1 | Current review process | 30.6 | | 31.3 | 38.1 |
| Condition 2 | Review plus tolerance | No tolerance – marks as submitted | 24.4 | 34.7 | 40.9 |
| | | Tolerance = 2 | 1.6 | 91.2 | 7.3 |
| | | Tolerance = 3 | 1.6 | 95 | 3.4 |
| Condition 3 | Single clean re-mark | 32.7 | | 23.3 | 44.06 |
| Condition 4 | Double clean re-mark with resolution | 36.9 | | 25 | 38.1 |

**Table B3: Summary of mark changes (unit C)**

| Condition | | Largest negative mark change | Largest positive mark change | Mean mark change (later mark − original mark) | SD |
|---|---|---|---|---|---|
| Baseline | Actual EAR process | −9 | 15 | 0.91 | 2.9 |
| Condition 1 | Current review process | −8 | 15 | 0.85 | 2.62 |
| Condition 2 | Review plus tolerance model (marks as submitted before tolerance applied) | −9 | 20 | 0.47 | 2.85 |
| Condition 3 | Single clean re-mark | −14 | 17 | 0.01 | 3.90 |
| Condition 4 | Double clean re-mark with resolution | −8 | 16 | 0.56 | 3.34 |

**Table B4: Percentages of mark changes (unit C)**

| Condition | | % scripts with negative mark change | | % no change | % positive change |
|---|---|---|---|---|---|
| Baseline | Actual EAR process | 18.8 | | 26.0 | 55.2 |
| Condition 1 | Current review process | 25.4 | | 21.4 | 53.2 |
| Condition 2 | Review plus tolerance | No tolerance – marks as submitted | 34.3 | 21.5 | 44.2 |
| | | Tolerance = 2 | 11.4 | 72.5 | 16.1 |
| | | Tolerance = 3 | 5.6 | 84.1 | 10.3 |
| Condition 3 | Single clean re-mark | 44.8 | | 12.0 | 43.2 |
| Condition 4 | Double clean re-mark with resolution | 36.8 | | 15.0 | 48.2 |

# Appendix C: Grade change charts for units B and C

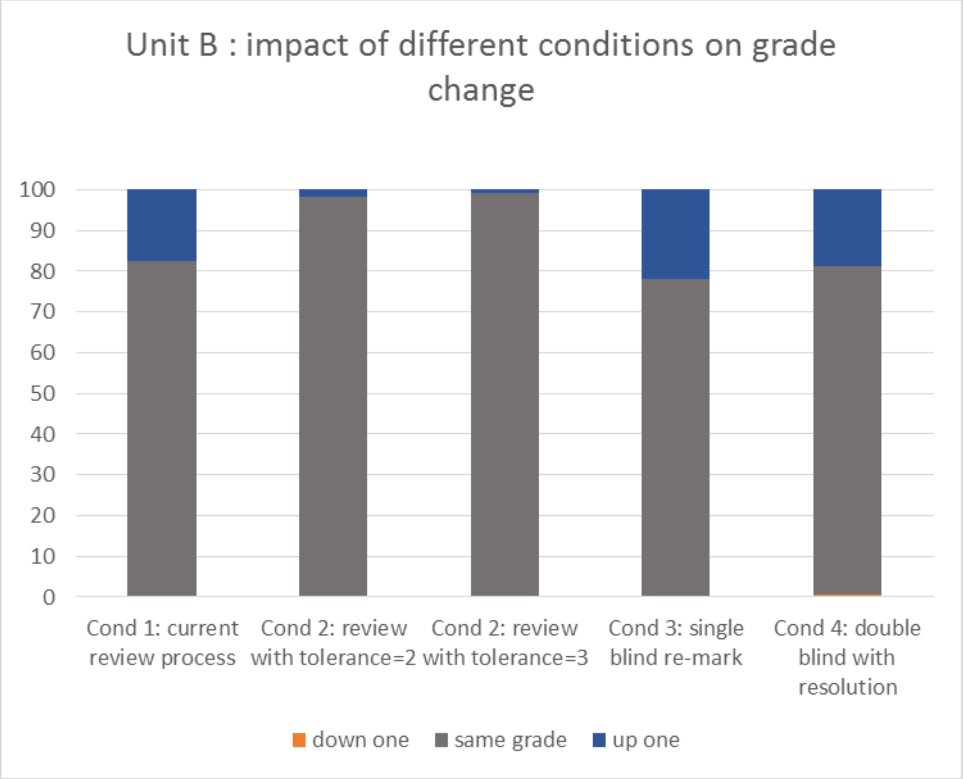


**Figure C1. Impact of different processes on qualification grade change (unit B)**

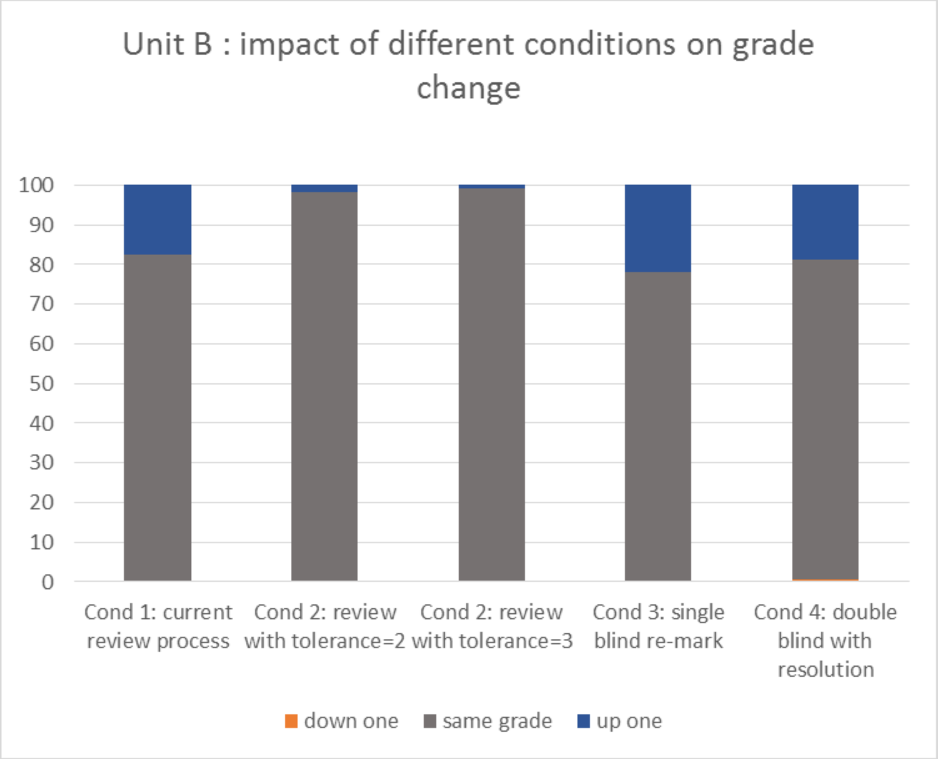


**Figure C2. Impact of different processes on qualification grade change (unit C)**

# Appendix D: Proximity to true score charts

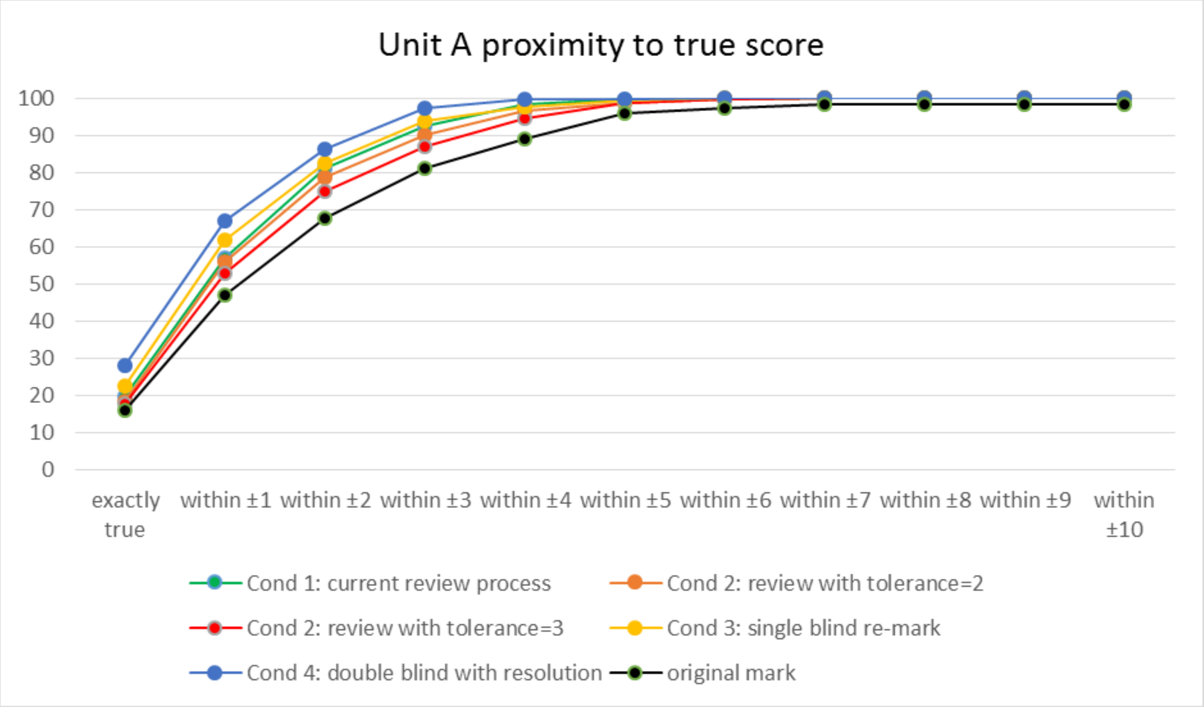


**Figure D1. Proximity to true score – cumulative percentage graph (unit A)**

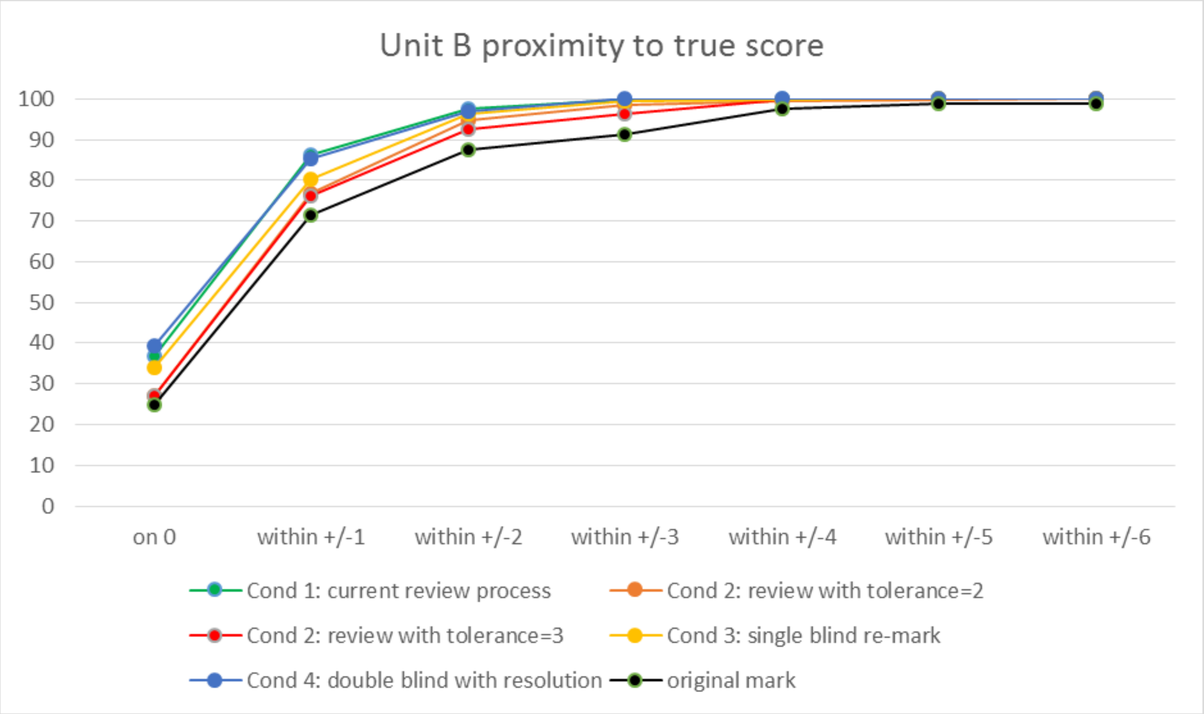


**Figure D2. Proximity to true score – cumulative percentage graph (unit B)**

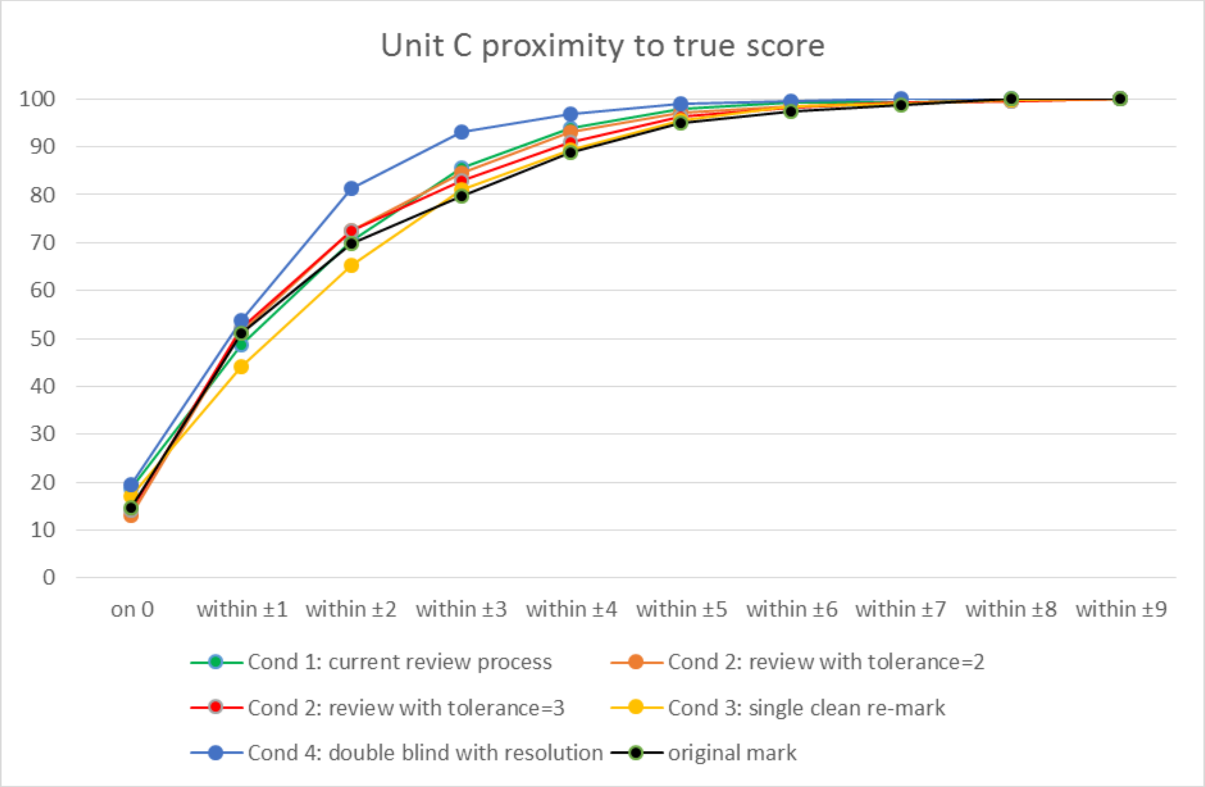


**Figure D3. Proximity to true score – cumulative percentage graph (unit C)**

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

| | |
|---|---|
| Telephone | 0300 303 3344 |
| Textphone | 0300 303 3345 |
| Helpline | 0300 303 3346 |