

Evaluation of Reviews of Marking and Moderation 2016

Study and survey



July 2017

Ofqual/17/6253

Authors

This report was written by Emma Howard and Beth Black from Ofqual's Strategy Risk and Research directorate.

Acknowledgements

The authors gratefully acknowledge the support of exam boards, reviewers and subject experts, as well as input and help from Emma Scott and Stuart Cadwallader.

Contents

Contents.....	3
Table of Figures.....	5
1 Executive summary	7
2 Background and introduction.....	9
3 Methodology – experimental review of marking study	13
3.1 Overview	13
3.2 Selection of subjects and units.....	13
3.2 Participants – subject experts and reviewers.....	15
3.3 Selection of Scripts	15
3.4 Unit materials	15
3.5 Tasks for reviewers and subject experts.....	16
3.5.1 Home based task for reviewers	16
3.5.2 Home based task for subject experts	16
3.5.3 Meeting tasks	18
3.6 A short note on the "definitive RoM mark"	19
4 Results for experimental review of marking study	19
4.1 Analysis	19
4.2 Script level analysis	21
4.3 Item level analysis.....	29
4.4 Qualitative findings.....	37
4.4.1 Mathematics	37
4.4.3 Biology	38
4.4.3 English literature.....	38
4.5 Response examples	41
5 Findings and discussion for experimental marking review study	46
6 Survey: introduction and research objectives.....	48
7 Methodology – review of marking and moderation surveys	48
7.1 Survey design	48
7.2 Participants	48
7.3 Respondent descriptives	49
7.4 Fieldwork	52
8 Results for the review of marking and moderation surveys	52

8.1	Reviews of marking survey.....	52
8.1.1	Guidance received for reviews of marking	52
8.1.2	Understanding the reviewers guiding principles to reviews of marking.	54
8.1.3	Making mark adjustments in reviews of marking	56
8.2	Reviews of moderation survey	63
8.2.1	Guidance received for reviews of moderation	63
8.2.2	Understanding the approach to reviews of moderation	64
8.2.3	Making mark adjustments in practice in reviews of moderation	67
8.2.4	The role of tolerance in reviews of moderation	68
9	Findings and conclusions – review of marking and moderation surveys.....	70
10	Overall conclusions and discussion	72
	References.....	75
	Appendices	76

Table of Figures

Figure 1. <i>Percentage of all GCSE reviews receiving each raw mark change in 2015 and 2016.</i>	10
Figure 2. <i>Percentage of all GCE reviews receiving each raw mark change in 2015 and 2016.</i>	11
Figure 3. <i>Percentage of GCSE reviews receiving each mark change by subject for 4 subjects.</i>	12
Figure 4. <i>Diagrammatic representation of the procedure for each set of subject experts and exam board reviewers.</i>	17
Figure 5. <i>The proportions of script mark changes by mark change type and subject.</i>	23
Figure 6. <i>The percentage of items with mark increases and decreases, by review session and subject.</i>	24
Figure 7. <i>Distribution of mark changes from the original mark as a percentage total script mark, overall and by subject, live review compared with experimental review.</i>	25
Figure 8: <i>Mark change difference between live review and experimental review, expressed as a percentage of the maximum script mark.</i>	26
Figure 9: <i>Tree diagram of mark change type at script level.</i>	28
Figure 10. <i>The proportions of item mark changes by mark change type and subject</i>	30
Figure 11. <i>The percentage of items with mark increases and decreases, by review session and subject.</i>	31
Figure 12. <i>Distribution of item mark change from original mark, overall and by subject, live review compared with experimental review.</i>	32
Figure 13. <i>Difference in item marks awarded at the live review of marking and the experimental review.</i>	34
Figure 14: <i>Tree diagram of mark change type at item level.</i>	36
Figure 15. <i>The number of marking reviewers by role.</i>	49
Figure 16. <i>The number of moderation reviewers by role.</i>	50
Figure 17. <i>The proportion of reviewers that conducted ROM in 2016 by the 20 most prevalent subjects.</i>	51
Figure 18. <i>The proportion of moderation reviewers that conducted reviews of moderation in 2016 by the 20 most prevalent subjects.</i>	51
Figure 19: <i>respondents' perceptions of feeling well-prepared and having a full understanding.</i>	53
Figure 20. <i>The number of reviewers that received training, by training type.</i>	54
Figure 21. <i>The distribution of responses indicating agreement with statements relating to understanding of the Review of Marking process.</i>	55
Figure 22: <i>Mean frequency of encountering particular marking errors when conducting reviews of marking. Bars show the standard deviation of responses.</i>	57
Figure 23. <i>The distribution of responses as to how likely it would be for the marking reviewers to change the mark in each circumstance.</i>	58

Figure 24. <i>The distribution of responses as to how much examiners agree to each points-based marking scenario.</i>	60
Figure 25. <i>The distribution of responses as to how much examiners agree to each levels-based marking scenario.</i>	61
Figure 26. <i>Responses indicating the time reviewers spent on ROM.</i>	62
Figure 27. <i>How marking reviewers' compare their own marking with the original marking.</i>	62
Figure 28. <i>How easy marking reviewer find it to understand how the original mark was justified.</i>	63
Figure 29. <i>Average scores for agreement in feeling prepared and understanding how to conduct reviews of moderation, by whether instructions and/or training was received.</i>	63
Figure 30. <i>The number of moderation reviewers by the type of training received.</i> ...	64
Figure 31. <i>The distribution of responses as to how important certain sources of information are in influencing judgements in reviews of moderation.</i>	66
Figure 32. <i>Distribution of responses regarding how moderation reviewers use the teachers' and moderators' marks and the candidate work to make judgements on making mark changes.</i>	68
Figure 33. <i>The distribution of responses to agreement to statements in how to deal with tolerance in reviews of moderation.</i>	69
Figure 34: <i>which mark to record – original moderator or original teacher mark – in different scenarios</i>	70

1 Executive summary

Data received from exam boards indicates that, despite changes to the rules around Reviews of Marking (formerly known as Enquiries about Results), there still remains quite a high number of small mark changes, including on assessments which are mostly or entirely 'subjectively' marked using 'levels based mark schemes'.

We conducted some research to understand the extent to which it was possible that such mark changes reflected Ofqual's rules; that marks should only be changed where the original mark represented a clear error in marking, or an error in the sense that it represented an unreasonable application of the mark scheme.

The research consisted of 2 strands:

1. An 'experimental' study involving reviewed scripts from units from 3 subjects representing a range of mark schemes from all boards for English literature GCSE, mathematics GCSE and biology AS level. In brief, the method involved senior examiners/reviewers and subject experts determining the 'definitive' mark to award at review if our rules were fully and properly implemented. We called this 'experimental review of marking' to contrast with the 'live review of marking'. This part of the research has given deeper insight into reasons for the marks given at review.
2. A survey of reviewing examiners and moderators (from all boards) to understand whether or not they had received training and/or instructions; and under what circumstances they would change a mark at review. There were around 1250 examiner reviewer respondents and around 180 moderator reviewer respondents. We estimate this to be between one third and one half of all reviewers.

Some key findings from the survey:

1. Only 46% of moderation reviewers and 64% of marking reviewers said they had received any training for undertaking review of marking/review of moderation training prior to undertaking reviews. The majority had received instructions.
2. Marking and moderation reviewers' responses appeared to show a good understanding of the principles of changing/not changing marks in line with our rules. However, when given specific scenarios, there was sometimes less consensus around whether or not they would or would not change the marks. For example, around 50% of marking reviewers indicated they would give a candidate 'benefit of doubt' at review when the original examiner had not done so. So it appears there might be a disjoint between principles and practice in some cases.

The experimental study has produced much data and key findings are outlined here:

1. In only 28% of scripts did the live review mark (R) agree exactly with the experimental review mark (E). In 60% of scripts, the marks from the live Review and Experimental review were different; and, in the remaining 12% of scripts, the examiners in the experiment were unable to jointly decide upon a 'definitive' review mark. These scripts were generally English literature scripts.
2. The scripts of particular interest are those 60% where the experimental review and the live review disagree. In one third of these cases, it appears that the live reviewer had mistakenly changed the mark where there was no error to correct. In around half of such scripts where experimental review and live review marks disagreed, both experimental and live reviewers had both changed the mark, and usually in the same direction, but that magnitude of the change was different. More often than not in these occasions, the original reviewers had changed the mark positively and of a greater magnitude than the experimental reviewers. However, there were some occasions whereby live reviewers had not changed the mark sufficiently. A diagram showing the complexity of the outcomes at script level is shown below in Figure 7 (a similar analysis has been conducted at item level as well (see Figure 14).
3. Some of the most interesting results are from the qualitative analyses of dialogue and reasoning between participants in determining mark allocations to particular items. It is clear that for English literature, a definitive mark or a definitive review mark is sometimes difficult to agree. This often appeared to be a result of different examiners evaluating responses which contained a mixture of relevant and irrelevant or wrong material. On some occasions examiners 'ignored' the irrelevant/wrong material and credited the rest. On others, the irrelevant/wrong material was deemed too significant and, as a result, undermined the perceived quality of the remainder of the response and/or the response as a whole.
4. Even in mathematics, there were rare occasions when the experimental reviewers disagreed with the live reviewers. On several occasions, it was because a candidate had used an unusual method in order to answer a more complex question that had not been recognised by either the original marker or reviewer. Sometimes in these cases, the 'layout' or presentation of the working may also have hampered examiners' attempts to fairly determine a mark.

2 Background and introduction

In the summer 2016 exam series, Ofqual put in place new rules relating to Reviews of Marking and Moderation (ROMM; previously known as Enquiries about Results). These rules mean that marks should only be changed at review if an error has occurred. These rules¹ were introduced because of concerns that sometimes small mark changes had been occurring in Enquiries about Results which were not correcting error, but rather replacing one legitimate mark with another legitimate mark², and often with a positive bias. Such a practice could give candidates/teachers who are willing and able to pay for this post-results service an unfair advantage. Also, there were some concerns that some marking errors in this system were not being amended at Enquiries about Results. The new Ofqual rules and accompanying guidance³ are clear on what constitutes error. In summary, error includes: (a) an administrative error, such as not adding up question totals correctly, or mis-entering a mark; (b) failure to apply the mark scheme and (c) an unreasonable exercise of academic judgement. Our new rules also required exam boards to provide appropriate training for reviewers (ie those examiners and moderators conducting reviews of marking and moderation). These changes were consulted upon and had come out of the research that we had conducted (Ofqual, 2015).

If reviewers conduct their reviews in line with the rules and guidance, this might mean that there are proportionately fewer mark and grade changes than in previous years as reviewers should not be changing one legitimate mark for another legitimate mark. The Official Statistics for Reviews of Marking (Ofqual, 2016) do indicate a very slight decrease, such that in 2016 18% of all AS/A level and GCSE qualification grades that were challenged were changed, slightly lower than in 2015 (19%). In 2016, 0.9% of GCE and GCSE qualification grades were changed, representing the lowest figure since 2013. If you were just to consider these headline

¹ These are laid out in Ofqual qualification conditions documents for GCSE and A level. See, for example, paragraph GCSE 17.4

<https://www.gov.uk/government/publications/gcse-9-to-1-qualification-level-conditions>

² For more information see : <https://www.gov.uk/government/consultations/markings-reviews-appeals-grade-boundaries-and-code-of-practice>

³ <https://www.gov.uk/government/publications/gcse-a-to-g-qualification-level-guidance#history>

figures of grade changes, while not always straightforward to interpret⁴, do not in themselves give strong support that the review of marking and moderation had been conducted very differently from previous years.

Looking at more granular data, ie item mark changes may also help evaluate the extent to which it appears reviews of marking and moderation have been conducted differently from previous years. Figure 1 and Figure 2 below are replicated from Ofqual (2017a). These figures show that for both GCSE and GCE the proportion of scripts with no mark change have increased, and the proportion of scripts with small mark changes have decreased, compared to 2015.

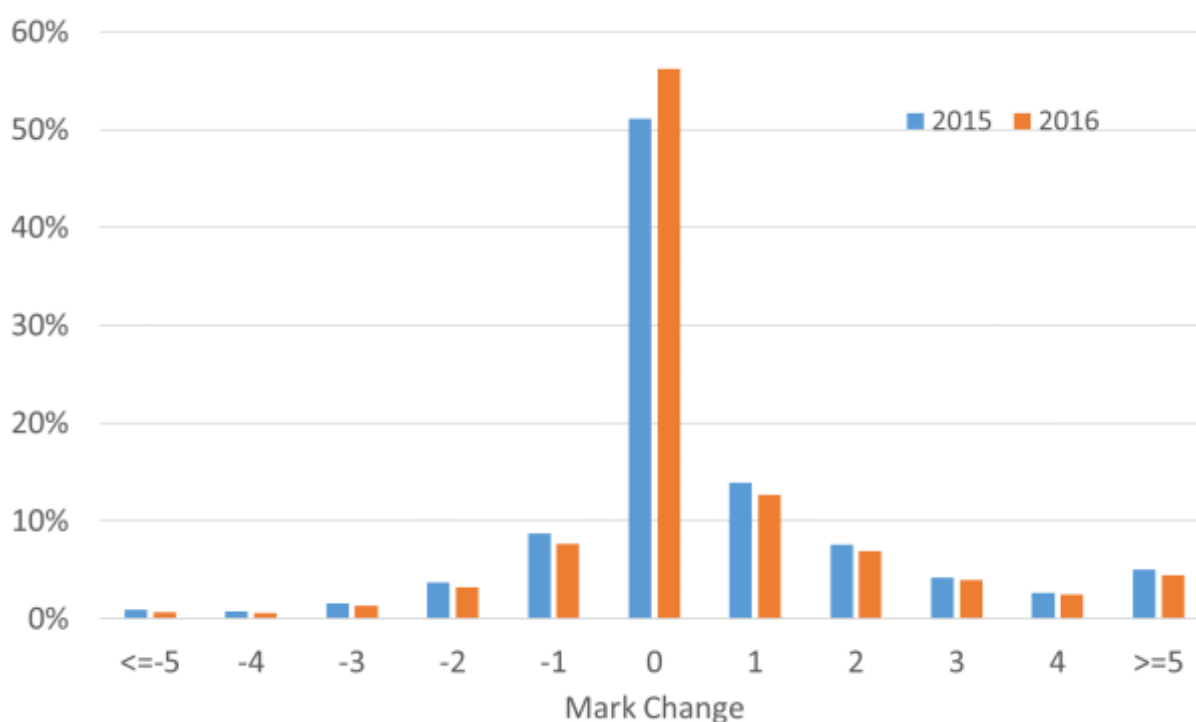


Figure 1. Percentage of all GCSE reviews receiving each raw mark change in 2015 and 2016.

⁴ Many factors have an influence on the proportion of grade changes e.g. the nature of the scripts put in for review, the proximity to the grade boundary, as well as reviewers propensity to change the mark either through correcting error and/or through substitution of one legitimate mark for another legitimate mark.

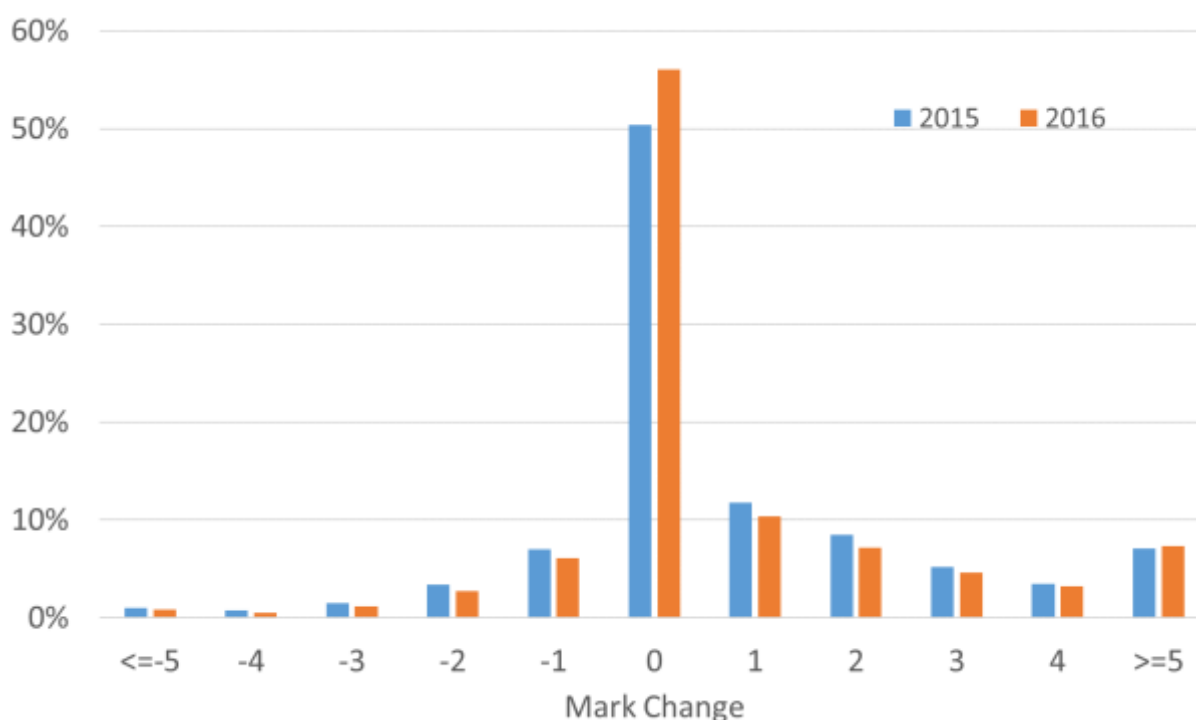


Figure 2. Percentage of all GCE reviews receiving each raw mark change in 2015 and 2016.

Across both GCSE and AS/A level, in 2016 nearly 30% (29.5%) of grades challenged through the ROMM process resulted in a mark change of up to ± 2 marks; and nearly 20% (19.4%) resulted in a mark change of ± 1 mark. If reviewers were reviewing scripts in line with new rules, this would mean that all of these one and two mark changes represent the correction of errors rather than substituting one legitimate mark for another. In some subjects, those which are more ‘objectively marked’, a one mark change is likely to represent a correction of an error; whereas in those subjects which are more ‘subjectively marked’, a one mark change is unlikely to represent a correction of an error.

In Figure 3, we can see the mark change distribution for four GCSE subjects, comparing 2015 and 2016. These four subjects represent 2 ‘subjectively marked’ subjects (English literature and English language) and 2 more ‘objectively marked’ subjects (mathematics and biology). We can see that in English language and English literature, there is a difference between 2015 and 2016, with fewer small mark changes in 2016. This could suggest the new rules have had some effect upon reviewers’ behaviours, that in 2016 they are less likely to substitute one legitimate mark for another legitimate mark. However, there does remain some small mark changes, particularly in English language. For mathematics and biology, we can see very similar patterns of mark changes between 2016 and 2015, indicating perhaps that these small mark changes do tend to represent correcting of error.

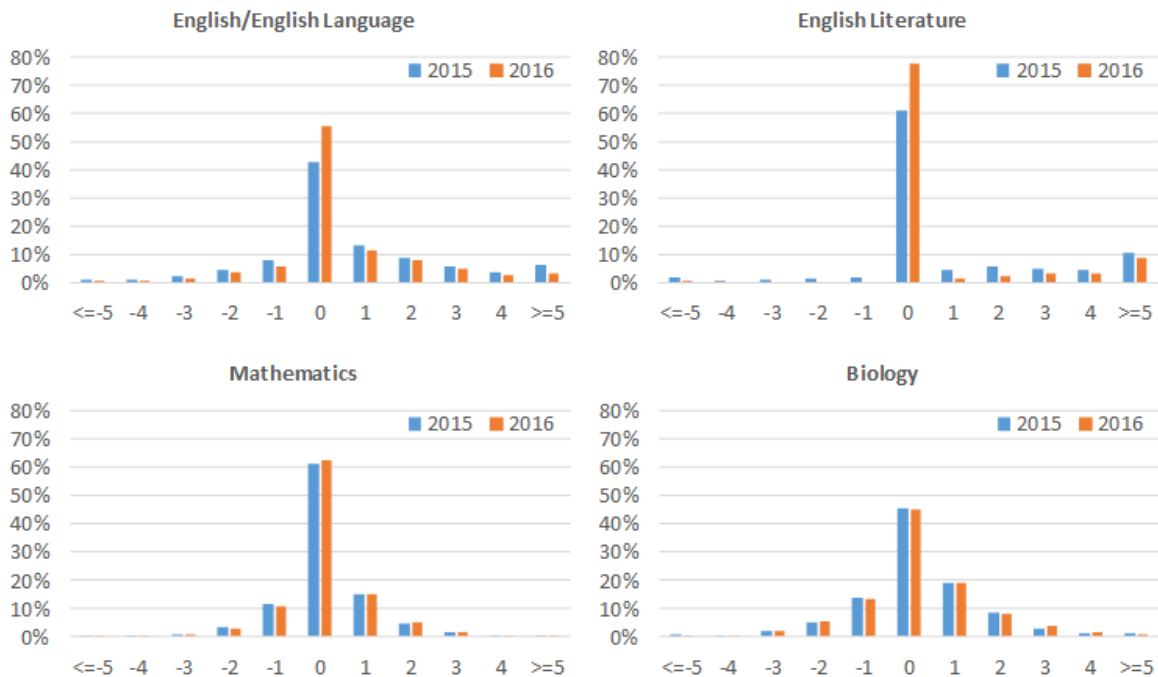


Figure 3. Percentage of GCSE reviews receiving each mark change by subject for 4 subjects.

Further inspection of mark changes in different subjects (Ofqual, 2017; Figure 21 and Figure 22) indicate broadly similar patterns, though noting that many 'subjectively' marked subjects (eg history, classical subjects) still have a not insignificant proportion of small mark changes.

Therefore, we wished to understand:

1. the extent to which mark changes made during reviews of marking reflect Ofqual's rules, ie to only correct error; and
2. the extent to which reviewers were trained to undertake reviews and understand the new rules and how to implement them.

So, we adopted 2 key data collection methodologies to help answer these questions:

1. An experimental study, involving reviewers and subject experts scrutinising scripts which had been through Reviews of Marking. Through a carefully structured process, a 'definitive review mark' (or 'definitive RoM mark') was arrived at, which could then be compared with the original (or 'live') review mark.
2. A survey of reviewers – both examiners and moderators from all exam boards, to understand whether or not they had received training in undertaking reviews and/or instructions, and to understand under what

circumstances they would and would not change marks at a review of marking or moderation.

We will describe the method and results for the experimental study (Section 3), followed by the method and results for the survey (Section 4).

3 Methodology – experimental review of marking study

3.1 Overview

The overall intention of the methodology was to arrive at a definitive RoM mark for each item and script, in a sample of scripts, from a small sample of units, and compare these with the marks generated in the live review of marking. Where there were differences between the 2 marks, possible reasons were gathered for this through inspection of items/scripts and analysis.

3.2 Selection of subjects and units

Three subjects were selected for this study: English literature GCSE, mathematics GCSE and biology AS level. These subjects were chosen on the following basis:

- to reflect a range of item and mark scheme types;
- to reflect a range of subject areas; and
- for having a medium or large volume of Reviews of Marking in 2016.

As a set of 3 subjects, it was hoped that this would give sufficient insight to understand the nature of mark changes at review and whether these reflected the rules to only change a mark where error was identified. Table 1 identifies the units which were involved in this study.

Table 1. *Units used in the study (figures rounded to the nearest 50)*

	Subject	Level	Unit	Number of ROM enquiries	Dominant mark scheme type
	Maths	GCSE	4365/2H	2,200	Points-based
AQA	Biology	AS	BIOL5	750	Points-based
	English literature	GCSE	97151H	8,550	Levels-based
OCR	Maths	GCSE	J567/03	550	Points-based

	Biology	AS	F215_1	500	Points-based
	English literature	GCSE	A663/02	750	Levels-based
Pearson	Maths	GCSE	1MAO/1H	19,700	Points-based
	English literature	GCSE	5ET1H/01	850	Levels-based
WJEC	English literature	GCSE	4202/02	3,000	Levels-based

These units have items with a range of maximum marks. As might be expected, mathematics GCSE has a high number of items, generally with low tariffs (most commonly out of 2 or 3 marks); biology AS level units have a similar profile; while English literature units have far fewer items per test, and generally with high mark tariffs (some questions out of 30 or 40 marks). Table 2 shows the range of item tariffs from the all the units in the study, by subject.

Table 2. *The number of items by the maximum item mark and subject.*

Number of items at each maximum mark			
Maximum item mark	Mathematics	Biology	English literature
1	18	34	0
2	48	28	0
3	35	13	1
4	10	10	4
5	8	0	0
6	1	1	0
7	0	0	1
8	0	0	1
10	0	0	2
12	0	0	1
16	0	0	1
20	0	0	2
25	0	1	0
30	0	0	2
40	0	0	2

3.2 Participants – subject experts and reviewers

For each of the 3 subjects, 2 subject experts were recruited from Ofqual's subject expert pool. These subject experts all have relevant subject qualifications and experience as teachers and examiners in the subject.

For each subject, examiners who had been marking reviewers in 2016 were recruited via exam boards. For each unit identified (see 3.1), we aimed to recruit 2 examiners who had conducted reviews in the 2016 session and the boards facilitated this. In all, 17 reviewers were recruited. All were senior examiners and many were Principal or Assistant Principal Examiners. There were 2 reviewers that took part in each unit included in the study, except for AQA maths in which only one reviewer was recruited due to a smaller number of reviewers that conducted RoM in this unit and availability.

3.3 Selection of Scripts

For each unit in Table 1, we selected a range of scripts which had been through the Review of Marking process in the 2016 session. We intended to select 20 scripts per unit on the basis of the mark changes in the live review of marking as follows:

- one or two scripts which did not have a mark change;
- one script with the largest increase in marks;
- one script with the largest decrease in marks; and
- all other scripts should have a similar profile of mark changes (in both direction and magnitude) to that of the whole unit in the live review of marking.

In some cases, we did not get our intended script selection. On some occasions the chosen scripts were unavailable or were not sent to us and/or different scripts were sent instead.

3.4 Unit materials

For each unit, the exam board provided the following:

- two versions of each of the script in the sample:
 - The original, prime marked script
 - The review of marking version of the script;
- question paper and related materials (eg the 'insert');
- the mark scheme;

- the 10 practice/standardisation scripts⁵ which were used to train markers in the original session; and
- Review of Marking instructions.

3.5 Tasks for reviewers and subject experts

Reviewers and subject experts both had similar but slightly different home-based tasks prior to coming to a one-day meeting. The ultimate aim of the exercise was to derive, as far as possible, a 'definitive RoM mark' for each of the scripts and items within. This was achieved through a carefully structured process of considering the scripts and dialogue between the participants. A diagrammatic representation of the procedure is provided in Figure 4.

3.5.1 Home based task for reviewers

Reviewers were sent all the question paper and mark scheme materials, and the standardisation scripts, and asked to re-familiarise themselves with the marking of the unit. All the reviewers had previously prime marked several hundred scripts and conducted reviews of marking on many scripts in this unit. The standard RoM instructions from their board were included as a reminder of the process. They were also sent a set of 10 scripts (either Set A or Set B depending on which pair they had been allocated to). These scripts had only the original marks on and not the live RoM mark. This reflects what can be seen in the live Review of Marking, except that in most instances the review takes place on screen, rather than on paper. The reviewers were asked to conduct a review of marking on each of the 10 scripts in their set of scripts and record their marks. The set of instructions can be found in the appendices.

3.5.2 Home based task for subject experts

The subject experts had not previously been standardised or had marked on this unit in 2016, so their home-based tasks were necessarily different. They were sent the same materials as the reviewers (question paper, mark scheme and standardisation scripts), including a set of 10 scripts (either Set A or Set B depending on which pair they had been allocated to). Subject experts were asked first to try to understand how to apply the mark scheme by studying the annotated standardisation scripts. Then, on this basis, look at each of the 10 scripts item by item, and decide whether or not the mark awarded was a 'plausible' mark or not; and whether there were other possible plausible marks for the same response. By 'plausible mark' we indicated this would mean 'a legitimate application of the mark scheme to a legitimate

⁵ In some cases this was fewer – depending on the number of scripts which had been used in the subject/boards standardisation procedure in the summer.

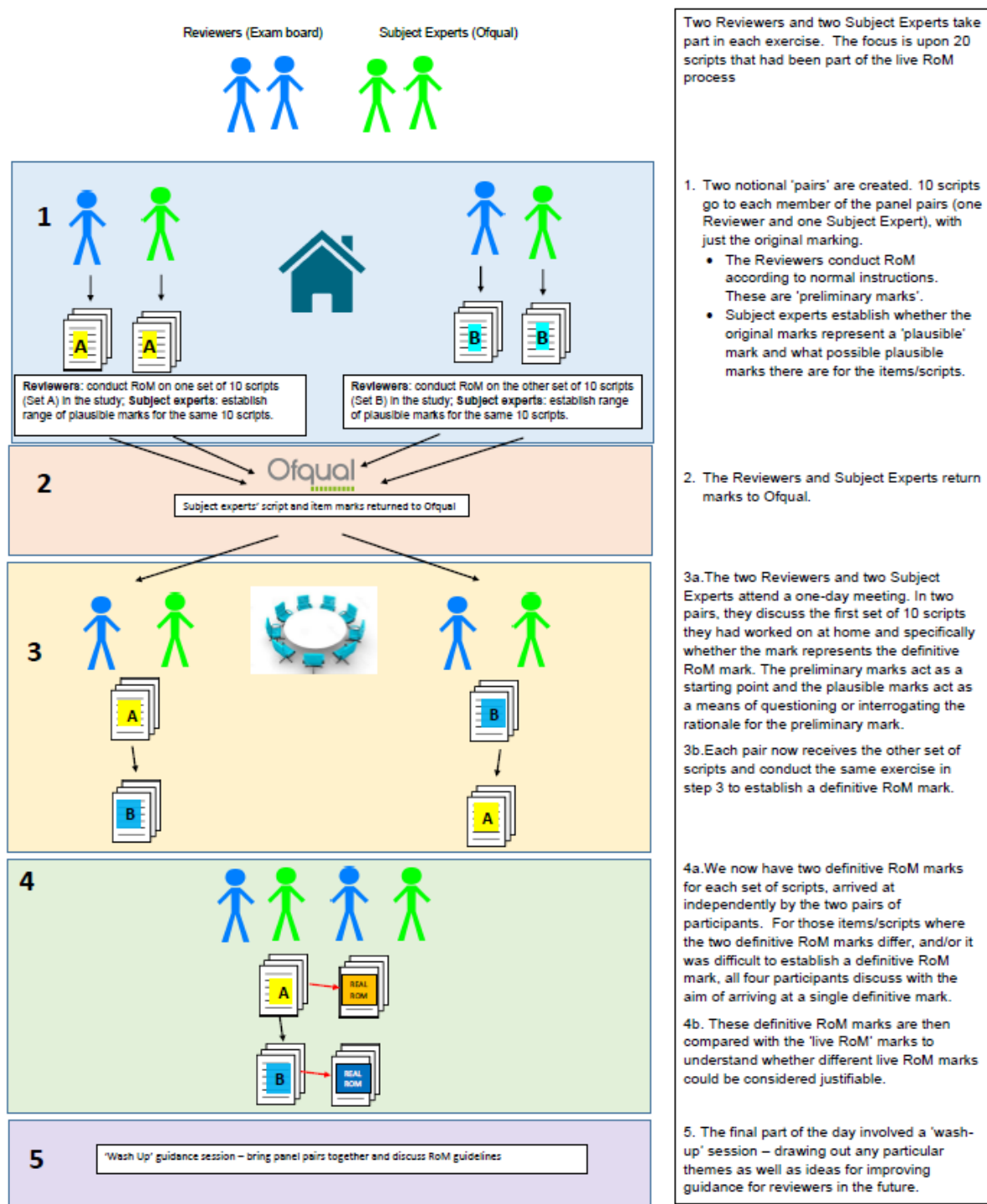


Figure 4. Diagrammatic representation of the procedure for each set of subject experts and exam board reviewers.

interpretation of the response'. Again, experts recorded these judgements, at item level, on a spreadsheet and returned them to Ofqual.

Instructions for the home based tasks can be found in the Appendices.

3.5.3 Meeting tasks

At the meeting, there were 2 reviewers and 2 subject experts. Each reviewer was paired with a subject expert so that there were 2 pairs. Two researchers also attended the meeting.

In the pairs, they were given the task of deriving a definitive RoM mark for each of the scripts they had considered as part of the home task. The researchers helped to facilitate the progress through the scripts as well as making notes about particular aspects of the discussion in arriving at the definitive RoM mark. To facilitate concentration and discussion, each pair had their own room.

The differing roles of reviewer and subject expert helped establish the definitive RoM mark. The reviewer had derived a 'preliminary' RoM mark. The subject expert, provided with the potential range of definitive marks for any item, could ask the reviewer how they had arrived at the preliminary mark and why they had or had not changed the mark and why one particular mark had been chosen over another. This meant there was a constant reminder around the rules to only changing marks where there had been an error. It facilitated an exposure of the thinking as to the nature of the error, and a rationale for the new mark chosen. This discussion occurred in a professional, collaborative, constructive and non-confrontational manner. During this process, there was regular reference to the standardisation/benchmark scripts. Together, the pair arrived at a jointly negotiated definitive mark. Where both agreed there should be no mark change, the discussion proceeded quite rapidly. Where the preliminary RoM mark did not coincide with the plausible marks, there was more discussion. Where there were difficulties in arriving at a definitive RoM mark, this was noted by the researcher.

Once definitive RoM marks had been established for the first set of scripts, they now proceeded onto the second set of scripts which neither had seen. The task was essentially the same, to establish a definitive RoM mark by way of the expertise of the reviewer, and the questioning and discussion triggered by the subject expert. If time was short, the pair was asked to prioritise particular scripts which were those which, unbeknownst to them, were ones which had been more problematic for the other pair. At the end of this phase, the items and scripts now had 2 'definitive RoM marks', one from each pair, as well as the original RoM mark.

In the final main phase of this study, both pairs came together in a single room, and discussed those scripts/items where the 2 sets of definitive marks did not match and/or had provoked discussion. The aim was to arrive at a definitive RoM mark agreed by all 4 participants. The group also discussed those instances where final definitive RoM marks did not agree with the live RoM mark.

To finish the day, the group had a final discussion over any recurring issues emerging from the items/scripts as well as implications for future guidance and training on conducting Reviews of Marking.

A diagrammatic representation of the procedure is provided in Figure 4.

3.6 A short note on the "definitive RoM mark"

We would argue that the process was likely to lead to a credible definitive RoM mark that could be legitimately used to evaluate the extent to which the live RoM marks were valid. The reasons for this are that:

- the process involved exam board reviewers working both independently and then together, to avoid biases;
- the process is akin to standard exam board processes for deriving definitive marks for practice, standardisation/qualification and seeding scripts in that multiple experts scrutinise and discuss responses and the merits of different possible marks. In many cases the reviewers in this study had also been involved in setting definitive marks in the live session;
- while the subject experts have not previously been standardised to mark on a unit, their role to continually check that changes to marks were an appropriate application of the rules for conducting RoM, to only change the mark in the case of error, worked well. It ensured an adherence to the rules, which is arguably less likely to happen consistently when reviewers are working alone, and without challenge and/or monitoring through inspection of script and item mark decisions.

However, we note that there are caveats. Some elements of the process might have affected marking judgements, such as the presence of Ofqual researchers, or the consideration of printed paper versions of scripts, rather than electronic.

4 Results for experimental review of marking study

4.1 Analysis

In all, data was captured for 175 scripts across all units and subjects in the experimental review of marking study.

The key method for analysis is comparing for (i) each script and for (ii) each item 3 marks:

O – the original mark which was the issued mark on results day in 2016;

R – the live Review mark, the mark which was issued as a result of the live Review of Marking;

E – the experimental Review mark, the 'definitive' RoM mark produced from the experimental study.

We were interested in the different patterns. For example, where the original mark, live mark and the experimental mark all agree ($O=R=E$), we could infer that the original marking was error free and that the 'live' Review of Marking was conducted appropriately. If the original mark was changed at both the live review and at the experimental review, and to the same mark ($O\neq R=E$), then we might infer that the original marking represented an error and that the live review of marking had properly amended the mark. A full description of these patterns is found in Table 3.

In a proportion of cases, the 4 participants could not agree upon a definitive RoM mark. These are designated as '?E'.

Table 3: *description of some main categories of items for mark different patterns of marks and inferences.*

Short-hand	Verbal description	Inference at item level
$O=R=E$	The original mark, live review mark and experimental mark all agree.	The original marking was error free and the live RoM was conducted appropriately.
$O\neq R=E$	The original mark was changed at both the live review and at the experimental review, and the live review and the experimental review marks agree.	The original marking represented an error. The live RoM properly amended the mark.
$O\neq R\neq E$	The live review and the experimental review both changed the original mark, but not to the same final mark.	The original marking represented an error, but the live review may not have amended the mark properly (eg did not change the mark enough or too much, or even in the wrong direction).
$O=E\neq R$	The experimental review agrees with the original marking. However, the live review had changed the original mark.	This indicates that the live review changed the original mark but that this was not correcting error, but substituting one legitimate mark for a different (legitimate) mark.
$O=R\neq E$	The live review of marking retained the original mark, but the experimental review changed the mark.	The live review did not correct marking error where it should have.

The analysis will describe script level outcomes and then the item level outcomes. The relationship between script level outcomes and item level outcomes is worth noting. Not all item mark changes result in a script mark change. Consider the following 2 scenarios:

- In a script, one item only has a mark change at RoM. This means there will also be a script mark change; and
- In a script, 2 items (or more) have a mark change at RoM. Where these mark changes are in different directions (one adding marks, the other subtracting), on occasion, they might cancel out, meaning that there might be no whole script mark change, despite there being item mark changes.

Importantly, in any script, it might be the case the vast majority of items were marked and reviewed correctly, but that just a single item mark change will change the whole script mark. Therefore, we might expect that item marks analysis will show high levels of agreement and unchanged marks; but script marks analysis will show lower levels of agreement and unchanged marks. This is illustrated in Table 4 which shows how, for example, 50% correct script marks relate to 90% correct item marks.

Table 4: Illustration of the relationship between item mark analysis and script mark analysis

	Script 1	Script 2	Overall
Question 1	Correct mark	Correct mark	
Question 2	Correct mark	Correct mark	
Question 3	Correct mark	Correct mark	
Question 4	Correct mark	Correct mark	
Question 5	Incorrect mark	Correct mark	
Item marking	4/5 correct item marking	5/5 correct item marking	9/10 (90%) correct item marks
Script marking	Incorrect	correct	1/2 (50%) correct script marks

For completeness, we describe script level results and then item level results.

4.2 Script level analysis

Of the 175 sampled scripts which went through the whole process, there were 9 scripts where all 3 marks agreed i.e. the original mark was unchanged by both the

live review and the experimental review (O=R=E). This indicates that these scripts were correctly marked and without error. Recall that deliberately disproportionately few scripts were selected which had no mark change at the actual review so this does not represent the proportion of error free marks which were issued.

A further 40 scripts had exactly the same mark change at live review and at experimental review. In these cases the live review of marking appeared to be conducted according to the rules, with error being appropriately corrected. Thus, the study indicates that in these 2 categories (represented by 2 shades of yellow in Figure 5), representing 28% of scripts, the RoM process was conducted properly.

The largest category of scripts, indicated by the turquoise colour, is where the live review and the experimental review changed the original mark, but not to the same mark total. This represents 38% of the scripts fully reviewed in the study. However, this is not a homogenous set of scripts: some mark changes at live review and experimental review are in the same direction, while some are in different directions. More description of the differences is given later (see description for Figure 9 for all subjects included in the study and Appendix B for descriptions by subject).

There were 35 scripts (representing 20% of scripts fully reviewed in the study) where the experimental review mark and the original mark agreed, indicating no error, but where the live review of marking had changed the mark. This indicates that the live review had likely changed marks where there was no marking error to correct.

There was also a small number of scripts (n=3; 1.7%) where the experimental review changed the mark but the live review had not, indicating the live review had not corrected error where it should have done.

The final main category is where the experimental study could not determine a definitive experimental review mark. This was the case for 22 scripts (12.6%) and indicates that some scripts, and some responses within, have marks which are not easily determined, usually because the response is of an unusual nature.

There are marked differences between subjects. English literature has the highest proportion of scripts for which it was not possible to determine a definitive experimental review mark; and mathematics is the subject which has the highest number of scripts where the live review mark agrees with the experimental review mark.

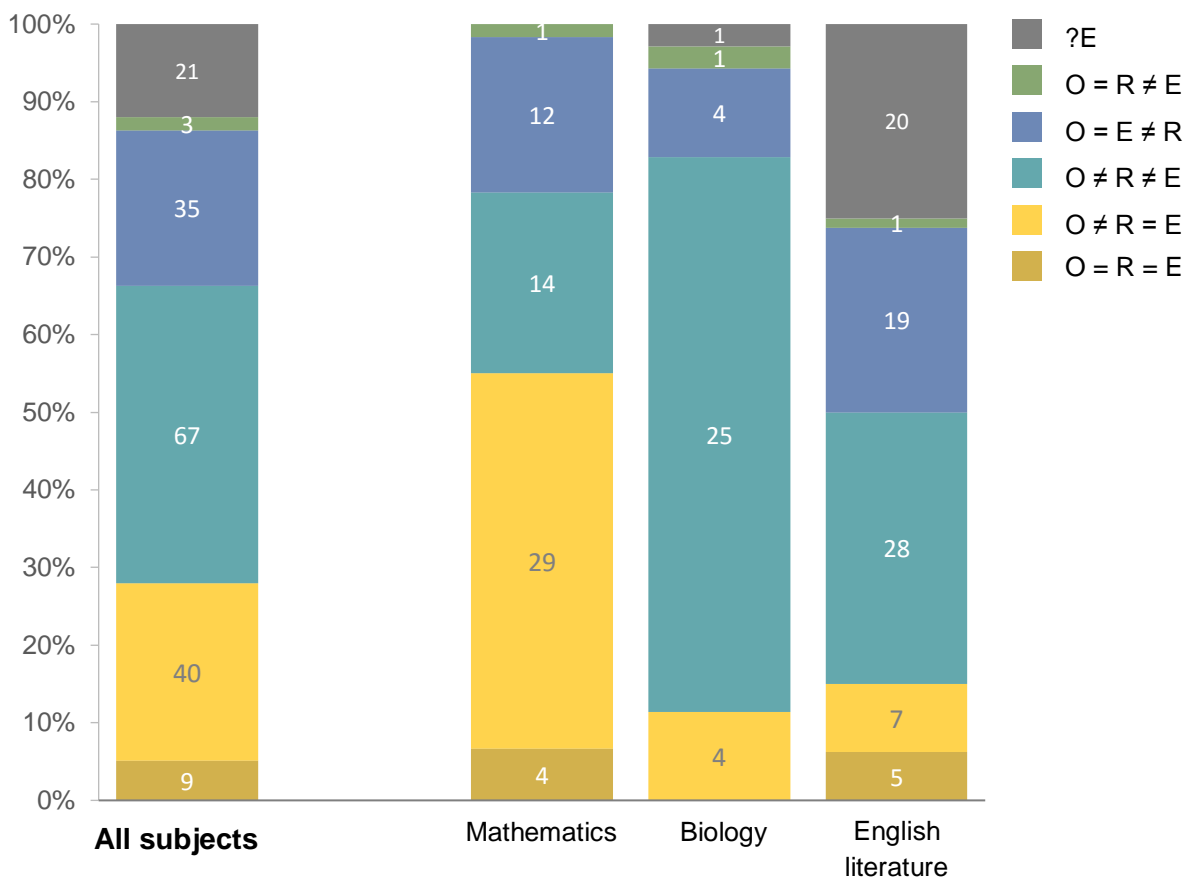


Figure 5. The proportions of script mark changes by mark change type and subject.

Note. Data labels represent the number of scripts

The direction of mark change, regardless of mark change type, are generally positive. Figure 6 shows the direction of mark changes (only for those scripts which had a mark change), for live review and experimental review, and broken down by subject. Generally, the number of scripts with mark increases is similar for both the live review data and the experimental review data. There are generally more mark increases than decreases (this is not surprising given that centres are likely to put in scripts/candidates where it is thought there is a greater likelihood of marks going up rather than going down). However, for biology, the pattern of experimental review mark changes were different from the live review - in the experimental review, there were more likely to be decreases than increases; whereas the live review had more increases than decreases. This suggests that the live review process in biology may have been biased towards finding candidates some additional marks and not taking off marks.

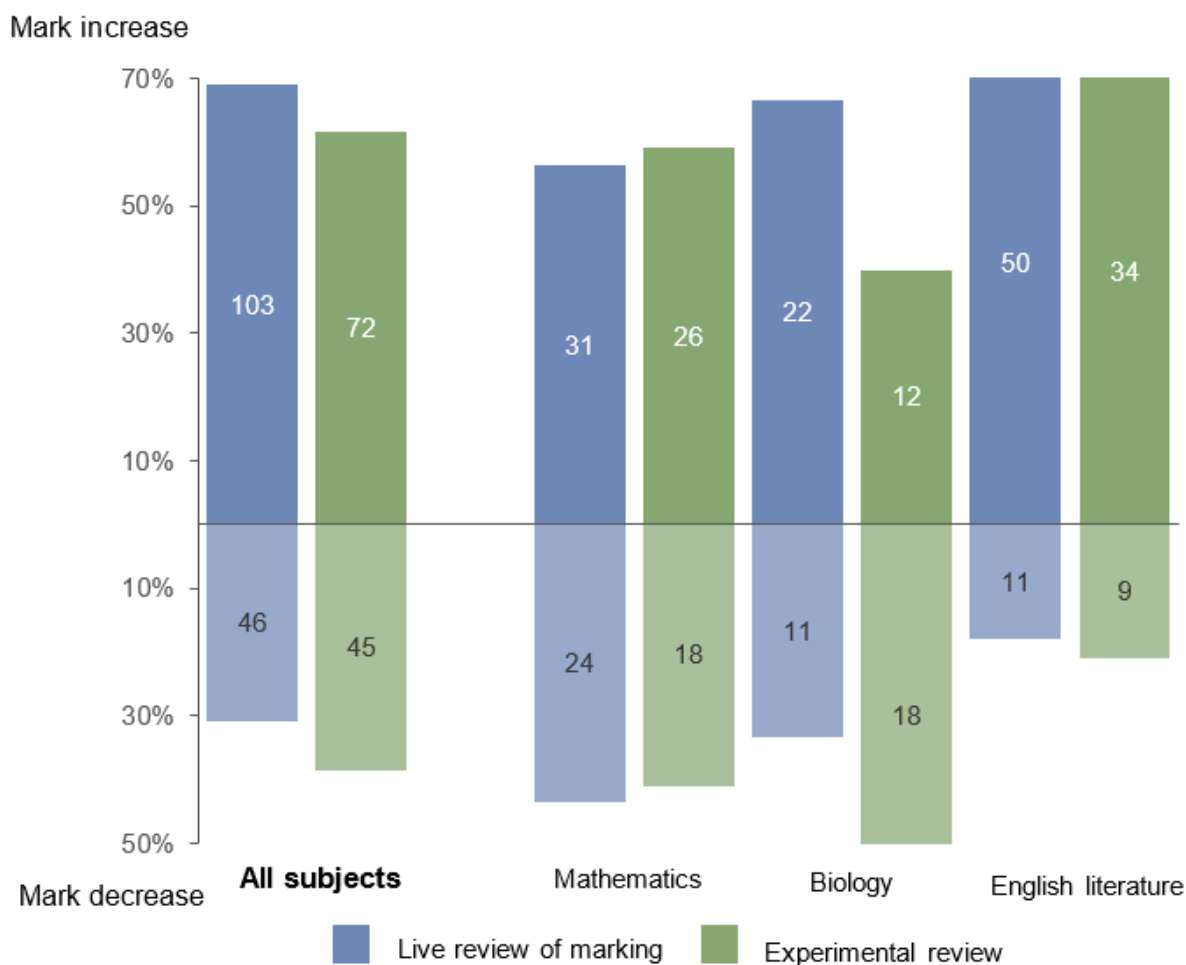


Figure 6. The percentage of items with mark increases and decreases, by review session and subject.

Note. Data labels represent the number of scripts; scripts with no mark change are not included, neither are scripts for which a definitive mark change direction at the experimental review was not determined.

Figure 7 shows the overall distribution of the mark changes at script level, comparing live review and experimental review, expressed as a percentage of total marks. The key point from these graphs is that there is a greater tendency for the experimental review to have a zero mark change compared to the live review. This can be seen where the peaks at zero are much higher for the experimental reviews (green) than for the live reviews (blue). This indicates that mark changes were made at live review which were unwarranted. This is not to say that reviewers acted in a way which was deliberate. Looking at the subject level distributions, we can see that this discrepancy is particularly apparent in English literature. The other key point is that the mark change magnitude is much greater for English literature than for mathematics and biology.

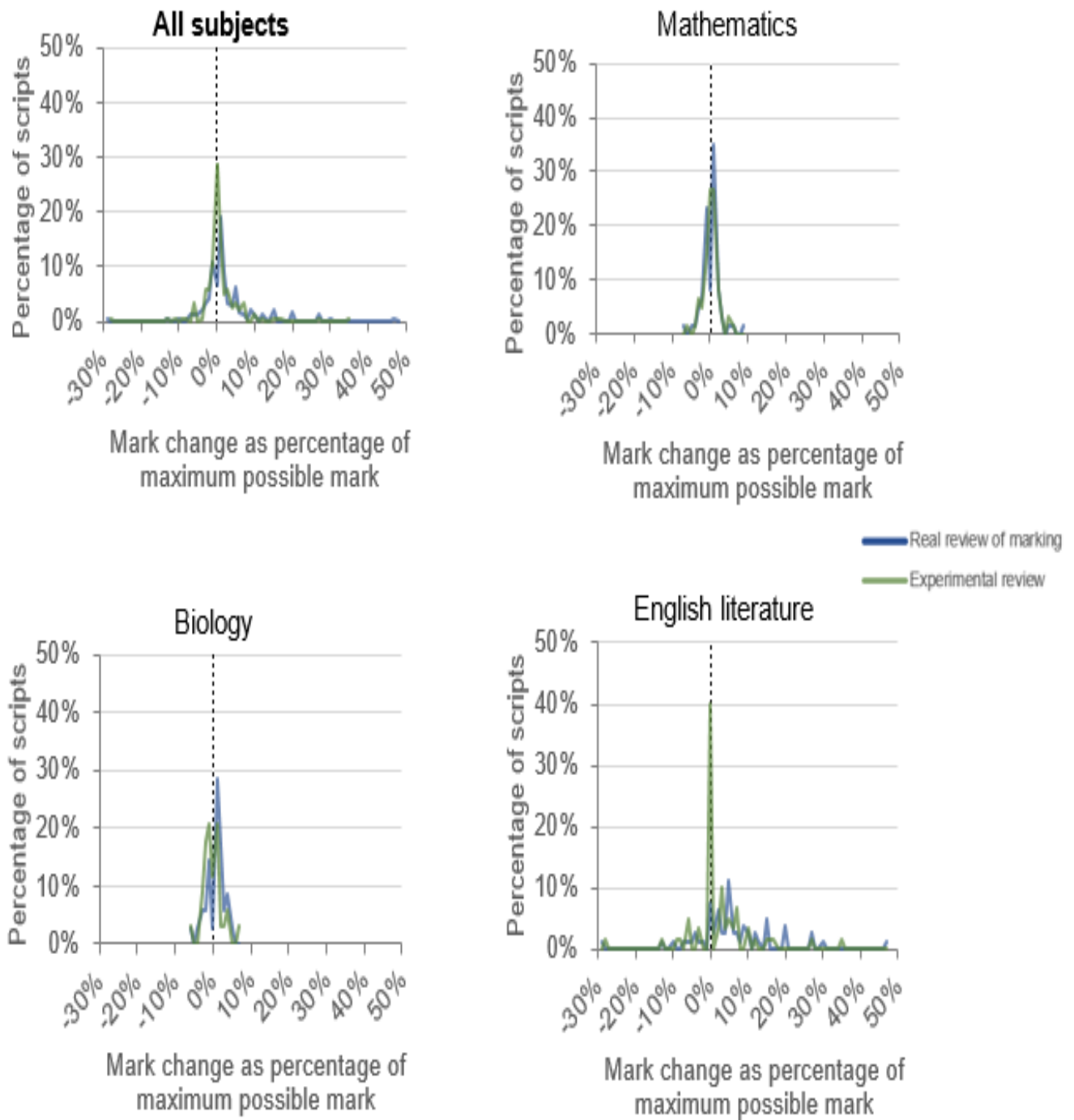


Figure 7. Distribution of mark changes from the original mark as a percentage total script mark, overall and by subject, live review compared with experimental review.

Figure 8 shows the mark disparities between the live review and the experimental review mark (or 'definitive RoM' mark). This is calculated, for each script, by finding the difference between the live review mark and the experimental review mark ('R-E'). A value of zero means that, the review and experimental review mark agree. A value of 2, for example, means that the live review marker gave 2 marks more than the 'definitive RoM' mark and so could be considered generous. A negative value, for example, -2, means that the reviewer was severe.

Overall, it can be seen that there are more scripts to the right of the line, indicating a positive bias in reviews of marking. The majority of scripts in the 'all subjects' graph show quite small percentage mark differences between the live and experimental review. This is true for both mathematics and biology where item tariffs tend to be quite low and the marking relatively objective. English literature, in contrast, shows quite substantial disparities between the live review and experimental review with some candidates receiving many more marks at live review than warranted according to the definitive RoM mark. One script over-benefitted by nearly 30%, and one script under-benefitted by nearly 10% of the overall question paper tariff.

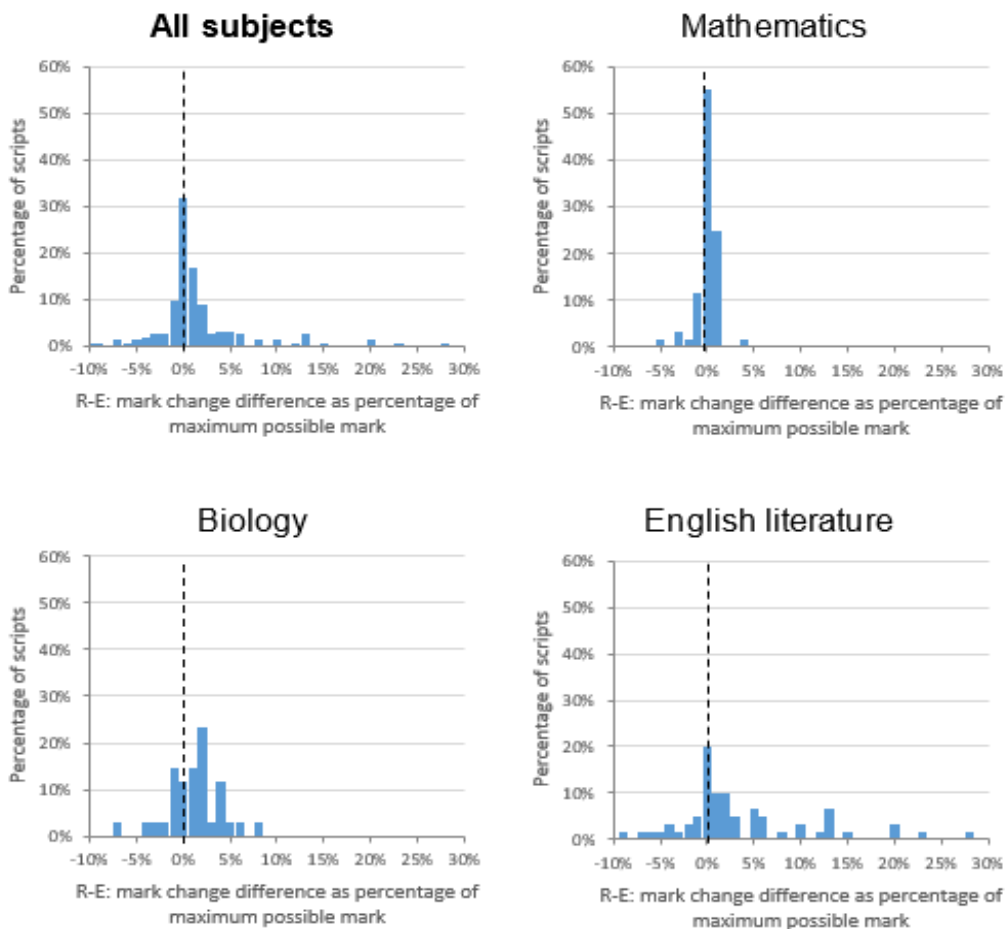


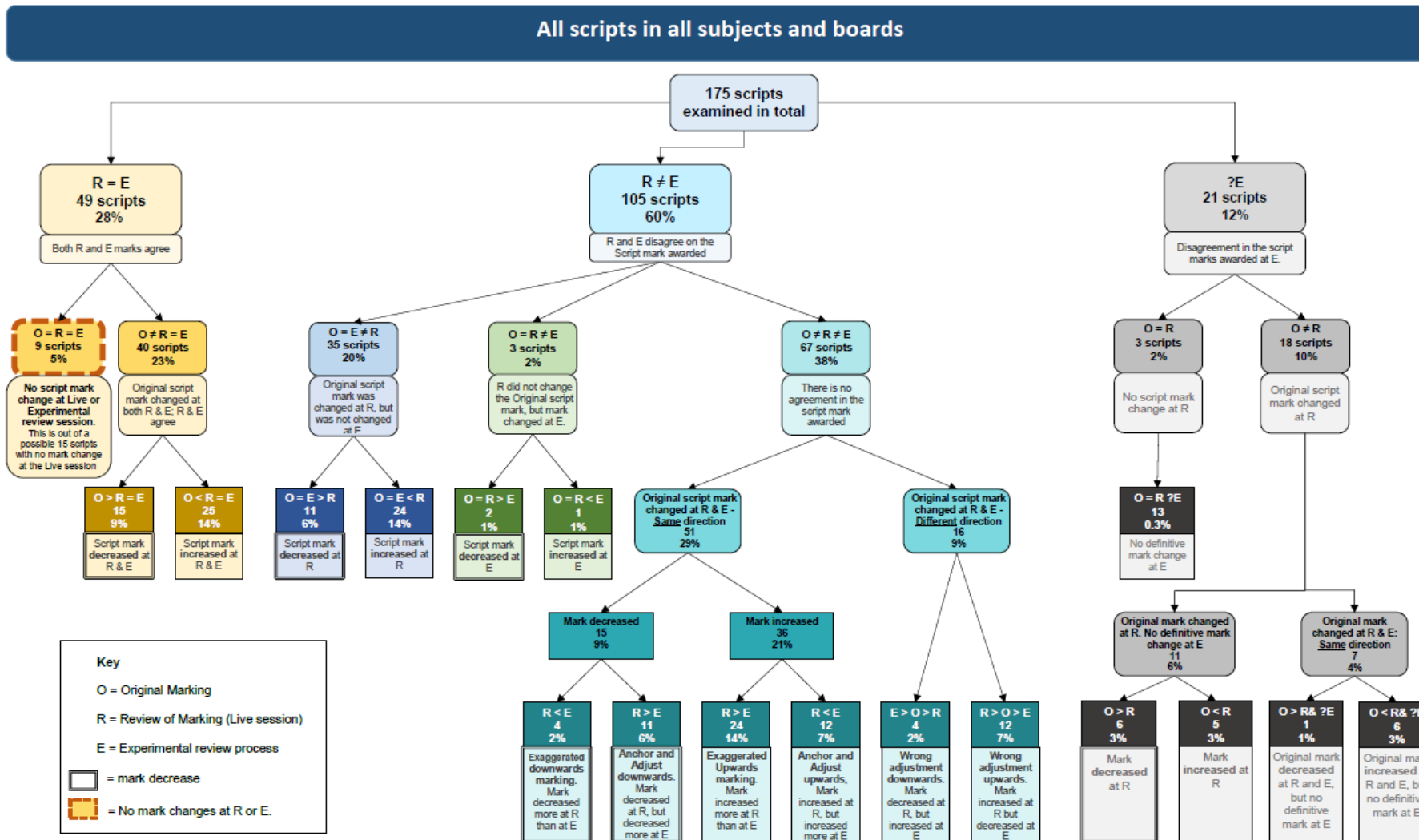
Figure 8: Mark change difference between live review and experimental review, expressed as a percentage of the maximum script mark.

To conclude this section of script analysis, we provide a ‘tree diagram’ of all the marking change types and subtypes. The headline figures are as follows:

- 28% of script marks at live review (R) and experimental review (E) agreed;
- 60% of script marks at live review and experimental review disagreed indicating that the live review had not been conducted in line with rules; and
- 12% of scripts had indeterminate experimental review marks – in other words, no definitive RoM mark could be identified. This was generally due to an interaction between an unusual response and the mark scheme; and most of these were in English literature.

Of most interest are those scripts where R and E disagree: there are various subcategories here. In the 20% of scripts where R had changed the mark, but E had not, the majority were in a positive direction – indicating that examiners had a positive bias when conducting a review of marking. This positive bias is also evident in 9% (n=16) of scripts where the experimental review and live review had changed the mark but in opposite directions. Twelve of these 16 scripts had marks added at live review which the definitive RoM indicated was not only unwarranted but that marks should have been subtracted.

Where R and E disagreed, albeit changing marks in the same direction (29% of scripts), overall there was a tendency for reviewers to add more marks than warranted, but take away fewer marks. There was also some evidence of ‘anchor and adjust’, that is, reviewers’ consideration of a new mark is conditioned by having seen the original mark and not wanting to go too far from this. Reviewers made more modest adjustments than the quality of the work was worthy of. In the majority of these cases, the anchoring effect was equivalent to 1 or 2 marks; but there was one instance of the experimental review taking the mark down a further 6 marks, and another of taking it up by a further 8 marks.



Note. Number indicates the number of items. % indicates the proportion of all scripts examined. Shaded out areas indicate instances in which mark adjustments were not made at the Review of Marking (Live session) and the Experimental review process. Percentages above 1% are rounded to no decimal places, percentages below 1% are rounded to 1 decimal place.

Figure 9: Tree diagram of mark change type at script level.

4.3 Item level analysis

We need more granular analysis of mark changes at item level in order to help us understand the script level outcomes. The response at item level is the basic unit of candidate performance and marker judgement. It is worth remembering that a script is made up of many items. A script may have, for instance, individual marks for 20 items, 19 of which are free from error and only one had an error in marking which was corrected at review of marking. So, we are likely to see proportionately few item mark changes but proportionately more script changes. (See Table 4 for an illustration of this point).

Figure 10 is the item level version of Figure 5. This figure indicates that the vast majority of items (90%) were free from error in the original marking. The next largest category (3%) of mark change type is where both the live review of marking and experimental review both agree on the mark change ($O \neq R = E$) indicating that the original review of marking was conducted appropriately. For just under 1% of items, it was not possible to determine a definitive RoM mark (the majority of these are English literature items). The remaining 6% of items in the study are in the various categories (see Table 3) where the review mark and the experimental review mark (the 'definitive' ROM mark) differ. This indicates that, at item level, 6% of items in the sample may not have been reviewed appropriately at the live review and candidates received different marks from the definitive RoM mark.

As with script level mark change categories, there are notable subject differences. Mathematics, which has the majority of items in the study, have the highest proportion of items (98%) in the 2 yellow categories which indicate that the live review mark agreed with the definitive RoM mark ($O = R = E$ and $O \neq R = E$). In comparison, biology has the next highest percentage (92%), followed by English literature with just 50% in the same 2 categories.

English literature is also notable because of high percentages items for which the reviewers and subject experts could not determine a definitive RoM mark. This is undoubtedly a reflection of the subjective nature of marking in this subject, the length of responses, and the greater likelihood of unusual responses. More discussion of such items is provided later in the section on qualitative findings. English literature is also unusual because the second largest mark change type is ' $O = E \neq R$ '. This means that live reviewers changed item marks, but the definitive RoM mark indicates that the original marking should have remained. This appears to indicate that English Literature reviewers are more likely to want to substitute one legitimate mark for another (legitimate) mark.

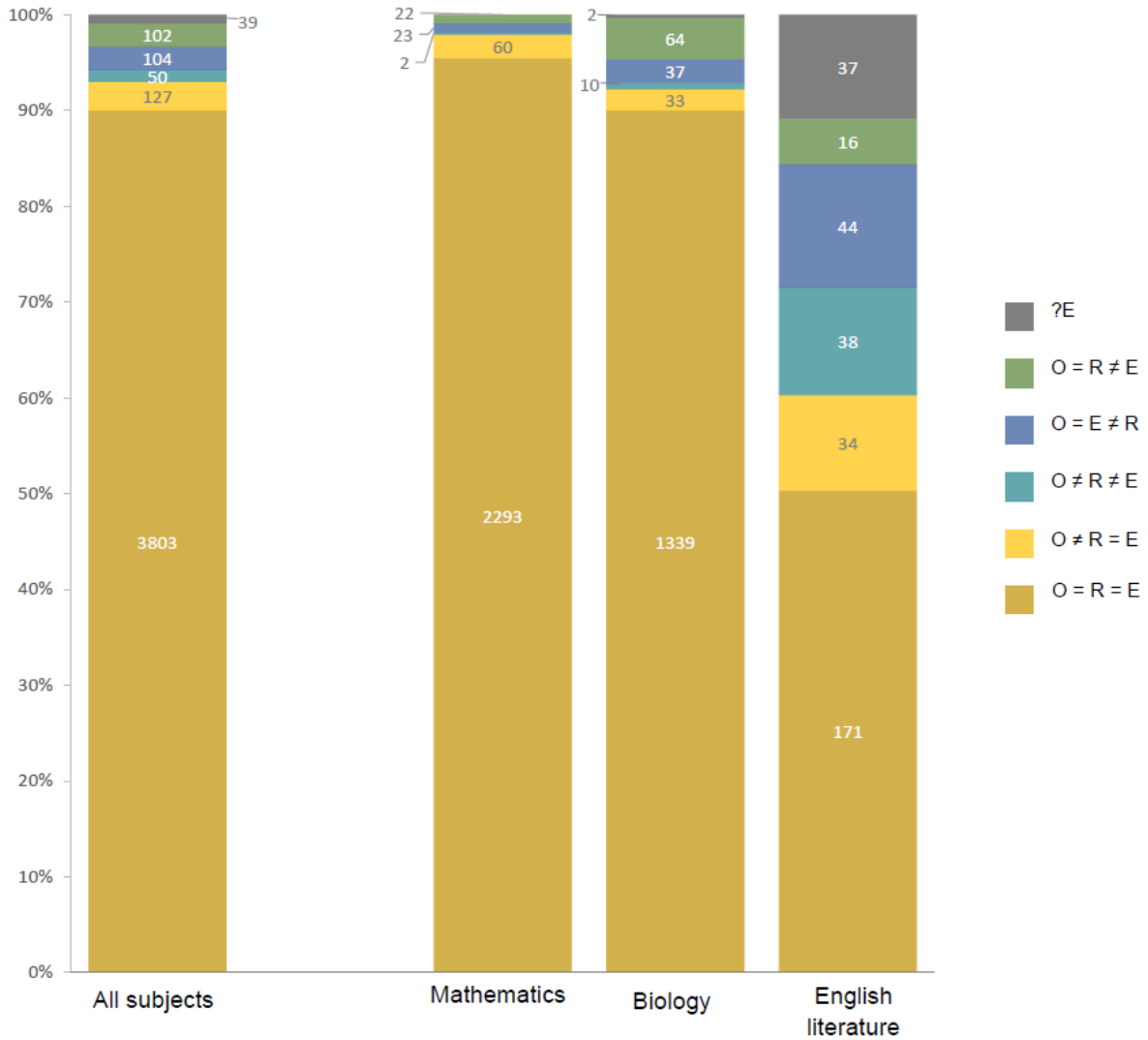


Figure 10. The proportions of item mark changes by mark change type and subject

Figure 11 is the equivalent item level chart to Figure 6 and shows whether mark changes were mark increases or mark decreases. A similar picture emerges whereby generally, the number of items with mark increases is similar for both the live review data and the experimental review data, and with overall more mark increases than decreases. As with the script level analysis, biology shows a different pattern from the other 2 subjects. The experimental review mark changes were quite different to the live review, in the experimental review, mark changes were more likely to be decreases than increases; whereas the live review had more positive than negative mark changes. This suggests that the live review process in biology was biased towards finding candidates some additional marks and not taking off marks.

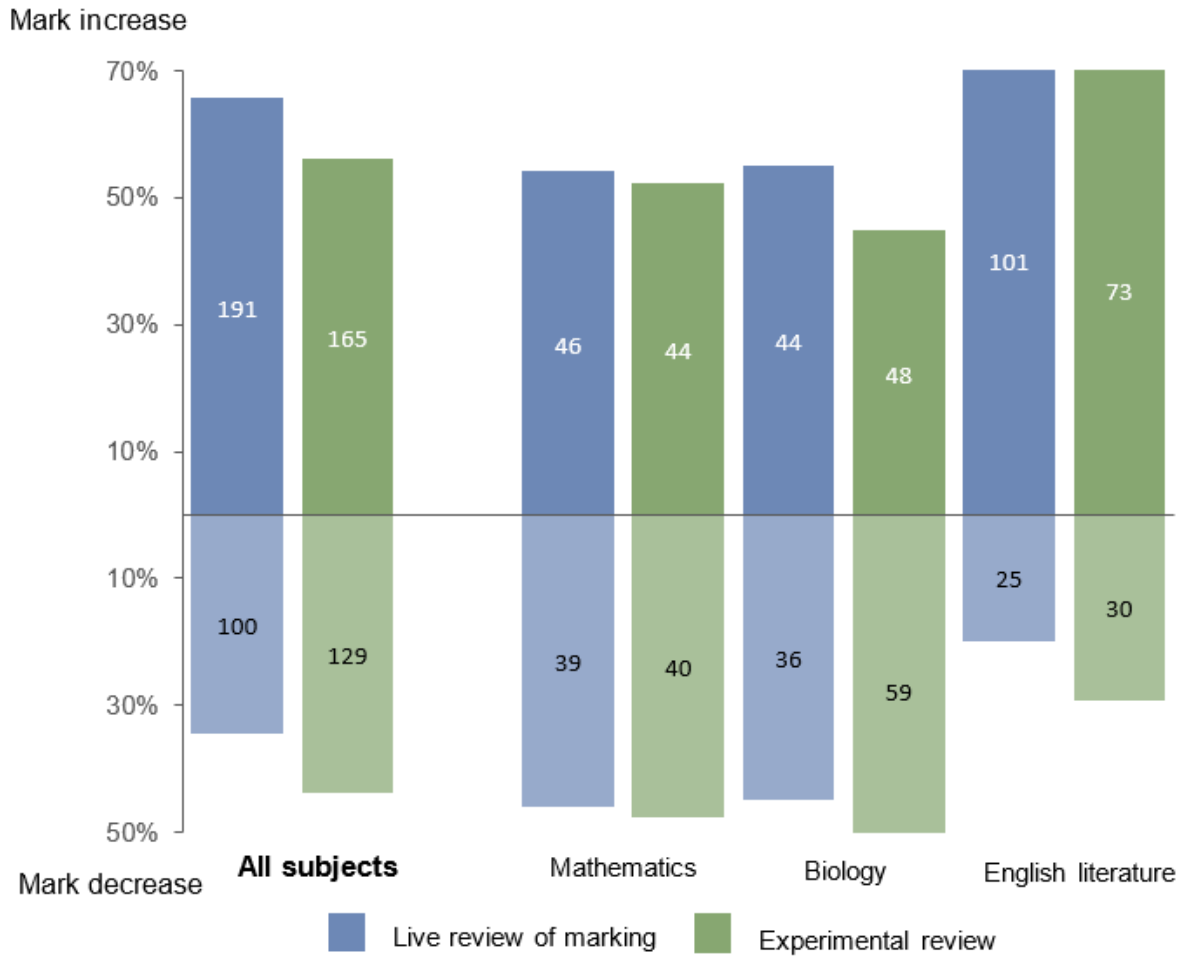


Figure 11. *The percentage of items with mark increases and decreases, by review session and subject.*

Note: items with no mark change are not included, neither are items for which a definitive mark could not be determined at experimental review. Data labels represent the number of items.

Below, in Figure 12, we can see the overall distribution of mark changes at item level, comparing live review and experimental review. The 2 sets of distributions look very similar, underlining the point that the vast majority of items in the study were marked correctly both at original marking and at the live review. This contrasts with the script level distributions (see Figure 7) where there are far more discernible and pronounced differences. The only real discernible difference between the distribution produced by the 2 processes is in English literature, where there are more positive changes in the live review of marking than in the experimental review.

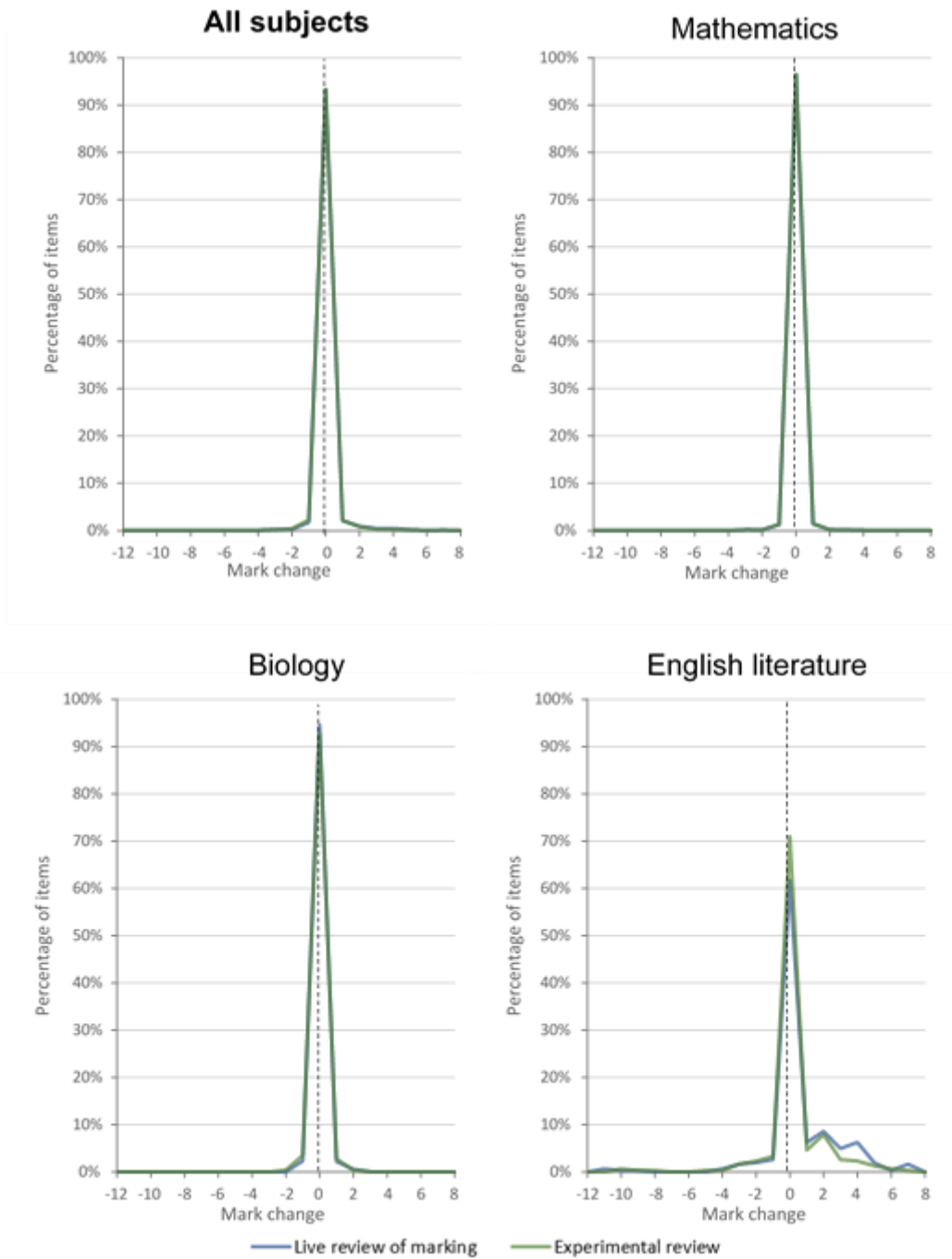


Figure 12. Distribution of item mark change from original mark, overall and by subject, live review compared with experimental review.

Note. Negative mark changes indicate where an item has been given a lower mark than at the original marking. Positive mark changes indicate where an item has been given a higher mark than at the original marking. Items where a definitive mark could not be awarded at the Experimental review (?E) have been excluded. The total number of items are as follows: all subjects: 4186; mathematics: 2400; biology: 1483; English literature: 303.

Figure 13 displays the disparities between the live review and the experimental review ('definitive RoM' mark). For each item, the difference between the live review mark (R) and the experimental review mark (E) is calculated. This figure is the item level equivalent to Figure 8. In general, most item mark disparities are 0 – in other words, the live review and the experimental review were in accord. Where there are mark changes, they are necessarily very small for subjects/papers with low item tariffs (see Table 2). However, there were some substantial mark differences on both biology (one where the original marking and the live review had over-marked an essay by 8 marks compared to the definitive RoM mark) and also in English literature where the live review of marking had in one case over-marked by 8, and another where it was under-marked by 6 marks.

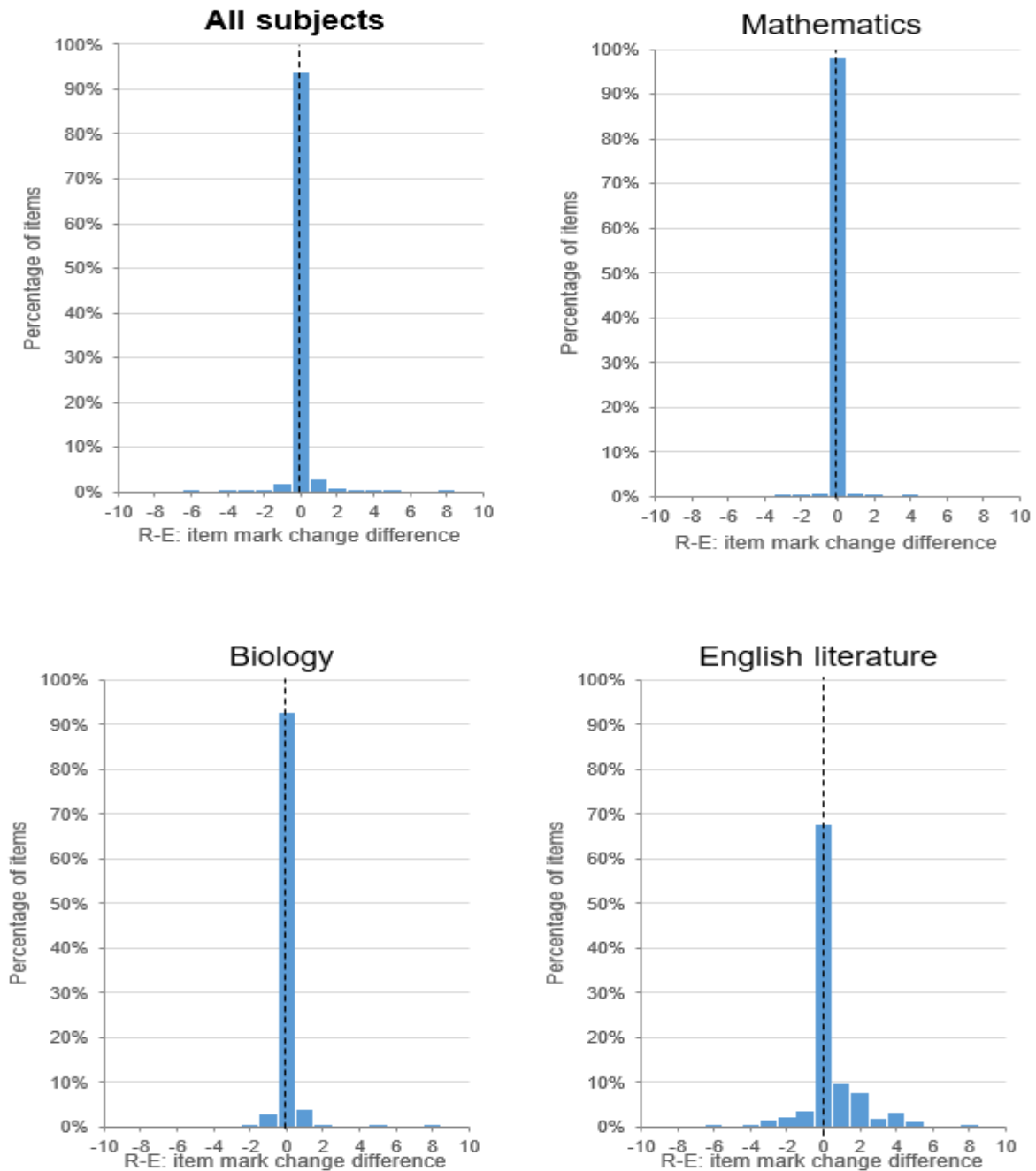


Figure 13. Difference in item marks awarded at the live review of marking and the experimental review.

Note. Difference in marks awarded at the live review of marking and experimental review are calculated as 'live review of marking mark minus experimental review mark'. Negative mark difference scores indicate where an item has been given a lower mark at the live review of marking than at the experimental review. Positive mark difference scores indicate where an item has been given a higher mark at the live review of marking than at the experimental review. Items where a definitive mark could not be awarded at the experimental review (?E) have been excluded.

To conclude this section of item analysis, we provide a 'tree diagram' of all the marking change types and subtypes (see Figure 14). The headline figures are as follows.

- 93% of item marks at live review and experimental review agreed;
- 6% of item marks at live review and experimental review disagreed indicating that the live review had not been conducted in line with rules; and
- 1% of items had indeterminate experimental review marks, in other words, no definitive RoM mark could be identified.

Most of the interest lies in those items where the experimental review and live review disagree. As with the script level analysis, generally, the live review and experimental review change marks in the same direction, and there is evidence of a positive bias in the live review.

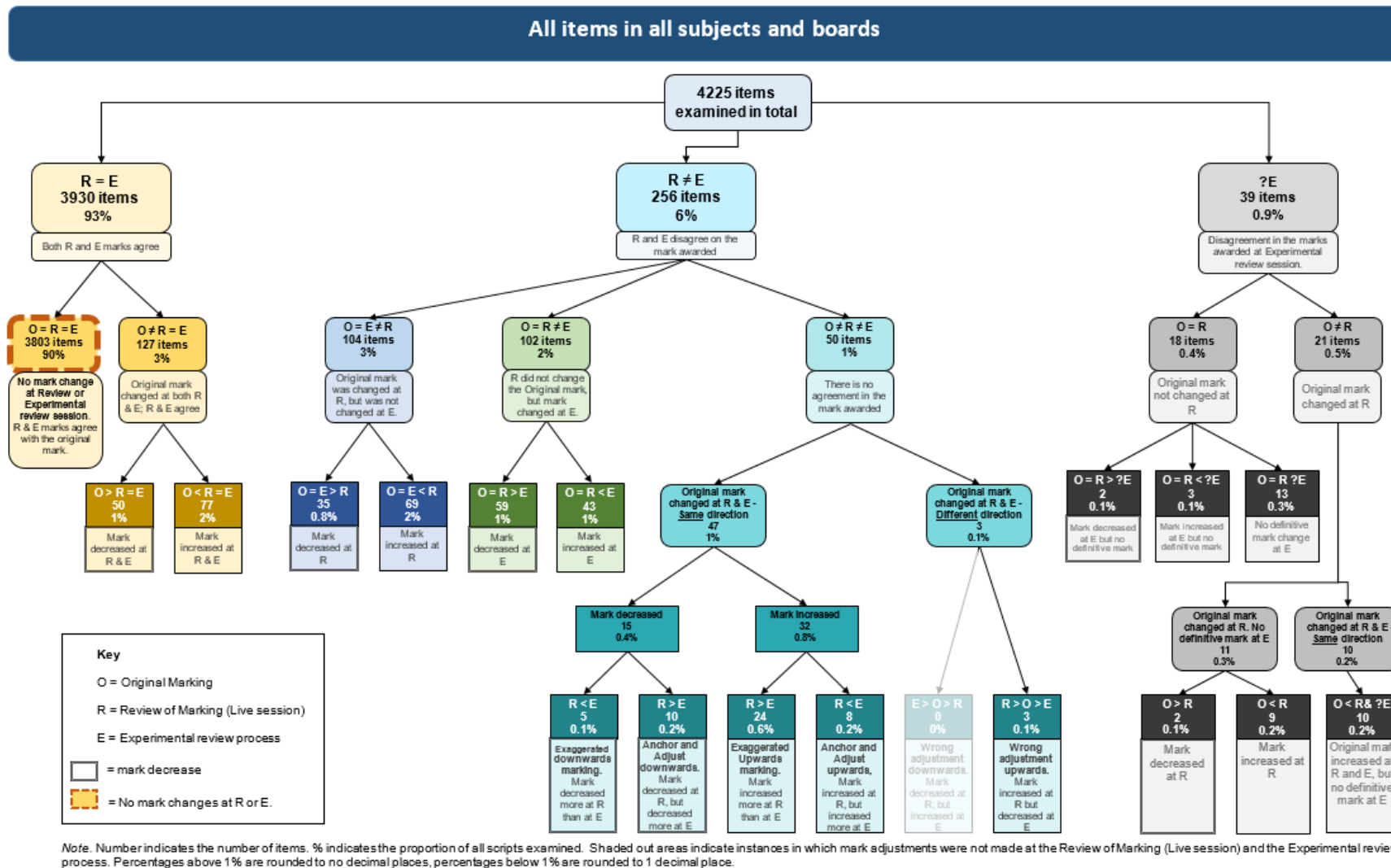


Figure 14: Tree diagram of mark change type at item level

4.4 Qualitative findings

Some interesting themes emerged from the discussions between the participants on the meeting day, which help to highlight sources of marker and reviewer error. It is not possible to capture all of the discussion around all of the items, or indeed quantify the regularity of particular recurring themes, but the main themes have been drawn out by subject and exemplified.

4.4.1 Mathematics

- **Marker error.** Where the experimental review changed marks in most cases this was straightforwardly identified as 'marker error' in that the original marker had missed a creditworthy point or had given a mark for a 'common incorrect response'. Reviewers tended to make comments such as 'they've just missed that bit there', implying some sort of attentional slip. Examples 1 and 3 provide illustration of these.
- **The importance of annotation.** The original annotation was very important for understanding how the original mark was arrived at – where the different types of marks (eg 'method' marks and 'accuracy' marks) were awarded. This was particularly important where candidates had partially attempted a question in more than one way and without arriving at a final correct answer.
- **Organisation of response.** Some responses which involved marker error, sometimes error at live review, were laid out in ways which made it difficult to follow the working and logic of the response. Furthermore, applying the 'marking rules' (with dependencies between different marks) to unusual as well as disorganised answers was quite complex and had sometimes caused marking error.
- **Alternative methods.** Candidates sometimes use 'alternative methods' to find a solution. In some cases in these scripts, these were rarely used alternative methods. On occasion, it appeared that the original markers, and in some cases, the live reviewing markers, had not spotted or recognised the creditworthy material. In some cases, the recognition of alternative methods may also have been hampered by presentation (see example 2).
- **There was a definitive RoM mark for all responses, but some took a lot of discussion.** In mathematics, there were some examples that were time-consuming in deriving a definitive RoM mark. These occurred particularly on those items which had higher mark tariffs, had more potential methods for correctly addressing the questions. They involved words and not just numbers, and had generally a broader and/or less well-specified 'outcome space'. A question's outcome space is the set of all (creditworthy) responses to it, actual and potential (see Marton and Saljo (1976) or Ahmed and Pollitt (2011)). In other words, where items are less 'constrained' or 'convergent' (ie having one or few fully correct possible answers) and therefore more 'divergent' (allowing

multiple possible correct answers, not all of which can be fully detailed in the mark scheme) reviewers encountered the greatest challenge.

4.4.3 Biology

- **Marker error.** Similar to mathematics, where experimental review changed marks from the original marking, this was usually straightforwardly identified as 'marker error', a marker had credited something not allowed by the mark scheme or had missed a creditworthy aspect. It was often attributed to an attentional slip. (See example 4.)
- **Candidates' language.** As with mathematics, the mark scheme and outcome space is generally tightly prescribed, and the marking is very 'rules-based'. However, in contrast to mathematics, candidates' use of words or abbreviations left more scope for interpretation and inference. For example, in one response a candidate's use of 'it', meant it was unclear whether they knew the correct answer or not; in another example, the candidate's use of the abbreviation 'BP' for blood pressure was discussed whether this was allowable or not.
- **Not reading the full response.** On some occasions, candidates had received credit in original marking which was removed at experimental review because the reviewers noted that while in the early part of the response the candidate had presented creditworthy material, they subsequently contradicted themselves or showed that they did not understand the material. The reviewers thought that original examiners had not read the full response. (See example 5.) It is probably that markers deploy 'scanning and matching' marking strategies (see Suto and Greatorex, 2008) for points-based mark schemes once they become familiar. This marking strategy, which means markers can quickly pick out the creditworthy material by looking for key words, may mitigate against reading the whole response and checking the overall understanding demonstrated. So when a candidate contradicts themselves, the examiners may not read or process this additional information at all because of the selectivity of reading engendered by the marking strategy. (See example 5, where the candidate contradicts the concept of stratified sampling by also talking about random sampling).

4.4.3 English literature

- **Original annotation.** The original annotation was very important for understanding the rationale for the original mark.
- **Standardisation scripts were very important.** Reviewers and subject experts made regular references to standardisation scripts. These were a source of comparison with responses under consideration. Reviewers would often say thing such as "this is better than script 5 which was a solid 'band 3'"

response and so it should receive x marks". On this basis, there were times where the reviewers declared original marks to be 'marker error'.

- **Word processed responses.** It might be expected that word-processed responses would be easier to judge on the whole as there would be no difficulties in interpreting handwriting. So, counter-intuitively, responses which were word processed were sometimes said to be more difficult to judge, mainly because it was difficult to interpret the length and hence substantiality of the response. There was sometimes a tendency for them to have been perceived as less substantial and therefore less detailed or 'considered', and therefore marks were often changed at review. As there are increasing numbers of scripts being word processed, marker perception and judgement of such scripts could be an important area of research.
- **Length of response.** Unusually short and good/particularly good responses were more challenging to judge. It was difficult to decide whether the candidates had shown enough evidence of particular aspects required in the mark scheme.
- **Dealing with wrong or irrelevant material.** Responses which contained one or more lengthy episodes that were wrong or irrelevant were also difficult to judge. There was often discussion on whether the passage(s) could be effectively ignored and just to credit that which was creditworthy; or whether the passage needed to be included as part of the overall consideration as it indicate 'lack of understanding' and so a 'holistic' judgement' rightly required this to be taken into account. Sometimes mark schemes did not help this decision in that they contained instructions or level descriptors which stated or implied both courses of action as appropriate, but which necessarily lead to different mark outcomes.
- **Tolerance.** Some boards apply a tolerance⁶ at item level during original marking for English literature, and which then becomes a 'rule of thumb' for Reviews of Marking and decisions whether or not to change marks. If a reviewer thinks the response is worth a slightly different mark but that this is within tolerance, then this mark should not or may not be changed. However, this did not appear to be a hard and fast rule. For example, sometimes, this rule of thumb, might be ignored if the response was deemed to have been put in the 'wrong level' of the mark scheme, or through a consideration across the rest of the script (see next point) it was decided that changing the mark could be justified. It is worth knowing that in Ofqual's rules, we do not require a tolerance to be operated during review of marking, mainly because operating a hard tolerance may prevent real error from being corrected.

⁶An 'allowed' mark difference, an examiner within tolerance continues to mark without intervention; but a mark outside of tolerance should trigger some sort of intervention eg a supervising marker discusses the marking error with them and/or they are temporarily suspended from marking.

- **Item consideration versus whole script consideration.** On occasion, whether or not to change a mark at review would be affected by interim judgements across all items in the script. For example, it was sometimes the case that both responses 1 and 2 were over-marked slightly, but not to the point of 'error', but when response 3 was also found to be slightly over-marked, there was a judgement that the mark to one or more responses should be moved to address the overall script mark. This is an interesting concept – that 'marking error' might exist at script level though not necessarily at item level.

4.5 Response examples

Example 1. A mathematics example of marker error which was corrected at the live review of marking.

(c) There were 16818 primary schools in England in 2012.

Work out an **estimate** of the mean number of pupils in each primary school.
Show clearly the rounded values you use.

16818 400,000 20,000

20,000 ÷ 400,000

(c) 400 [2]

Example 1 shows the response to a question on the estimation of the number of pupils in primary school.

- The original mark was 1.
- The live review of marking changed the original mark and awarded 0.
- The experimental review changed the original mark and awarded 0.

The original marker awarded one method mark for the presentation of '20,000 ÷ 4,000,000'. However, the candidate has written the latter value in the thousands instead of millions: '400,000' and so is incorrect and should not be awarded a mark. It is likely the original examiner made an attentional slip.

Example 2. A mathematics example of an alternative method used, which was not credited by the original marking or live review of marking

(b) Work out the probability that, on a day in June, it does **not** rain and my tennis match is cancelled.

The image shows handwritten mathematical work for the problem. On the left, the numbers 0.75 and 0.8 are written. Below them, a standard multiplication grid is shown but crossed out with diagonal lines. To the right of the grid is a vertical column of numbers: 7, 14, 21, 28, 35, 42, 49, 56. To the right of this column, the text 'So or 1' is written above '50 2', with '(2)' written below it.

Example 2 shows the response to a question on the probability that it will not rain and a tennis match will be cancelled.

- The original mark was 0.
- The live review of marking did not change the original mark and awarded 0.
- The experimental review of marking changed the original mark to 1.

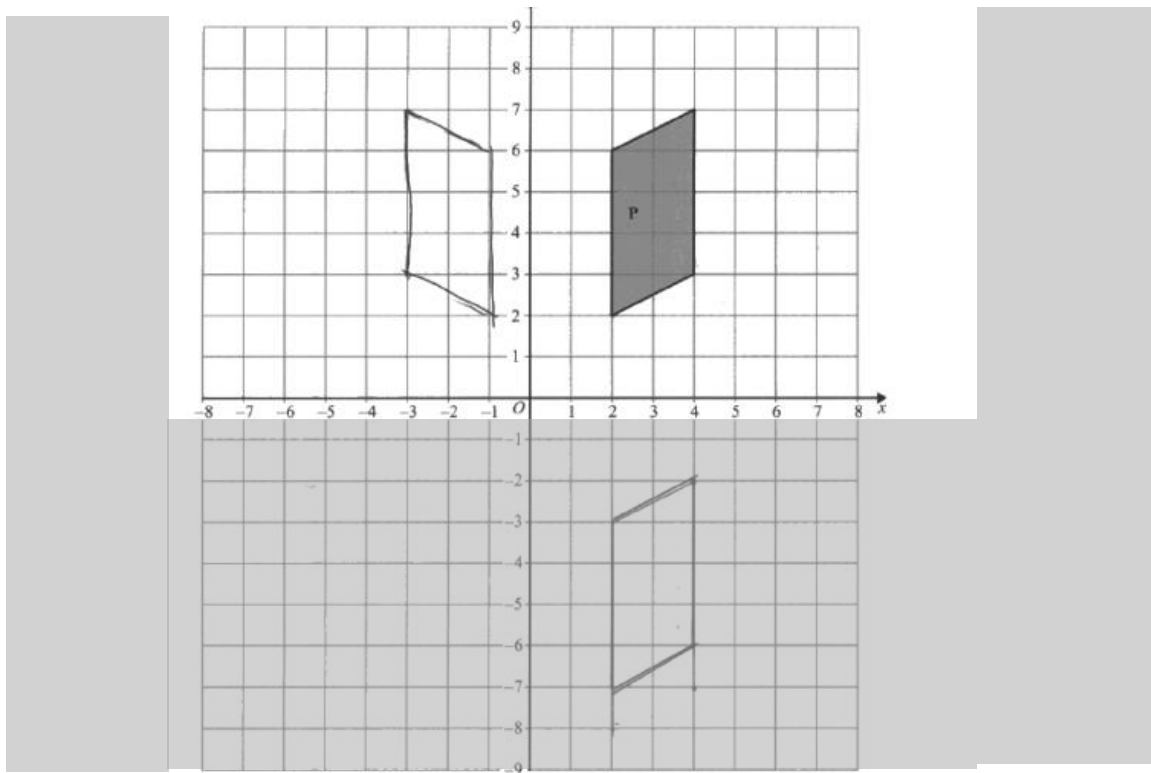
The mark scheme awards 1 mark for the method of multiplying 2 numbers (0.75 and 0.8), and 1 mark for the correct answer.

The experimental review highlighted the part of the responses on the left hand side of the answer area, which although appear to be crossed-out working that should be ignored, is an alternative multiplication method called a 'lattice multiplication'. This response can therefore be credited the method mark for the lattice multiplication working of 0.75×0.8 .

The original examiner and reviewer may have also been hampered in recognising the 0.75 and the 0.8 because the decimal points are very small and 0.8 has been presented vertically rather than horizontally.

Example 3. A mathematics example of where presentation of the scripts in the marking software may result in marker error.

This question requires the candidate to reflect a shape in the line x axis ($x=-1$).



- The original mark was 1.
- The live review of marking and experimental review changed the original mark and awarded 0.

The mark scheme awards partial credit (1 mark) for a reflection in a vertical or horizontal line, and awards the second mark for the reflected shape being on the correct co-ordinates.

The response shows a shape on the top left side which is a reflection of shape P in a vertical line, but in the wrong vertical line, so on its own would be worth 1 mark. However, the candidate has also drawn a second shape on the bottom right side, which is not crossed out and is incorrect (it is a translation rather than a reflection). The presence of this incorrect shape therefore negates the mark awarded for the first shape because the candidate has given 2 possible answers, one being incorrect.

It was thought that the presentation of the response in the marking software was the cause of the original marking error. The 'clip area' (the area immediately visible to markers) only showed the top part of the graph. Examiners would have to actively go outside the clip area to check the bottom half of the graph area. So this is a special type of slip, which the marking software has 'helped'. This highlights the care needed for specifying the clip areas.

Example 4. A biology example of original marker error which was an attentional slip.

(ii) The woodland in Fig. 1.2 can supply timber continuously, sustainably and economically.

Discuss some social, aesthetic and ethical benefits of managing woodland in this way compared to coniferous monoculture.

> Social benefits ~~is~~ is that some trees may provide medicinal value, and leads to a variety of ^{wild} species. Also managing woodland reduces the risk of succession.

> The aesthetic benefits is that ~~if~~ the woodlands would look ~~at~~ aesthetically pleasing leading to increased ~~at~~ tourism. Ethical benefits, is that ~~all~~ all species of organisms have the right to ~~live~~ live. [4]

Example 4 shows a response discussing benefits of a method of managing woodland:

- the original mark was 1; and
- the live review of marking and experimental review of marking changed the original mark and awarded 2.



The original marker awards a mark for the candidate identifying benefits for tourism. However, the mark scheme indicates that one mark can also be awarded to the candidate's response with regards to the woodlands being aesthetically pleasing.






Both of these benefits are introduced in the same sentence. It is therefore possible that the original marker missed the 'aesthetically pleasing' point.

Example 5. A biology response which was not fully read by the original marker or marking reviewer.

Researchers could investigate the distribution and abundance of grasses across the salt spray and rain-watered zones.

Choose and describe sampling techniques that would show how the distribution and abundance of native and introduced grasses change across the two zones.

  In your answer, you should make clear how the equipment is used.

 A quadrat  could be used ~~in~~ ^{random} places along the beach and any ~~living~~ organism touching one of the spikes could be counted. Once everything has been counted, a mean could be made in that zone before repeating  in the next. Instead of random, the sampling could be stratified so ~~that the~~ that the researcher could easily compare 5 square metres of the salt spray zone and 5 square metres of the rain-watered zone. These measurements could be made along a tape measure  at specific intervals  decided before the experiment. [7]

- The original mark was 4.
- The live review of marking changed the original mark and awarded 3.
- The experimental review of marking changed the original mark to 2.

The mark scheme awards 1 mark each for ‘using a quadrat’ and for ‘measuring of samples at set intervals’ (the first and last ticks in the example). The original marker has also awarded one mark for the sampling technique being ‘stratified’, however this is *contradicted* in the response with ‘a quadrat could be used [in] random places...’. The mark-scheme indicates to not credit ‘randomly’. At both the live review of marking and the experimental review, a mark was therefore removed for this contradiction. A further mark was removed at the experimental review because the reference to a ‘tape measure’ in the response does not meet the requirements of the mark scheme which indicates that the tape measure must be described as laying across the zones or in a line, which the response fails to do. It is therefore possible here that the original and reviewing examiners scanned the response and matched both ‘stratified’ and ‘tape measure’ to the mark scheme, awarding marks which either contradicted or not fully substantiated.

5 Findings and discussion for experimental marking review study

The majority of items in the study were error free from marking error, and the reviews of marking outcomes matched the definitive marks from the study. However, a proportion of item live review marks did not match the definitive item RoM marks (around 6%) and this translated into around 60% of scripts in the study having marks at live review which did not match the definitive script RoM mark.

There is some evidence of live review of marking having some positive bias. For instance, in some of these cases (29%), marks at live review and experimental review were in the same direction but differed in magnitude. There was a tendency for this mark to be most pronounced for positive changes, ie live reviewers gave more marks than experimental reviewers. Additionally, in 14% of scripts the live review of marking had changed the original mark upwards where the definitive RoM did not at all, suggesting that not only is there a positive bias, but that the reviews of marking were still replacing one legitimate mark with another, in other words changing marks where no marking error is present.

This, alongside the mark change data published⁷ suggests that exam boards implementation of the new rules around Reviews of Marking are partial rather than full. Reviewers are still changing marks where no marking error is present.

Some of the most interesting results are from the qualitative analyses. It is clear that for English literature, a definitive mark or a definitive review mark is sometimes difficult to agree. This often appeared to be a result of different examiners evaluating responses which contained a mixture of relevant and irrelevant or wrong material. On some occasions, examiners 'ignored' the irrelevant/wrong material and credited the rest. On other occasions, the irrelevant/wrong material was deemed too significant and, as a result, undermined the perceived quality of the remainder of the response and/or the response as a whole. These marking difficulties are not confined to conducting a Review of Marking – these are marking difficulties that are likely to have affected a proportion of scripts in original marking and reduce overall marking consistency. Given that English literature is associated with lower levels of marking consistency (Ofqual, 2016b), attempting clear guidance to markers on how to deal with such responses might be a worthwhile pursuit.

There is also evidence that on occasion, live reviewers left marking errors unamended. The apparently small number of scripts (n=3; 1.7%), where the

⁷ <https://www.gov.uk/government/publications/exam-and-assessment-marking-research>

experimental review changed the mark but the live review had not, could be an underestimation of this phenomenon. Only 15 scripts in the study had had no mark change at live review and so these represents 20% of such scripts in the study. This may indicate that some reviewers were confused about their task in that they did not understand that marking error should be corrected, no matter how small.

The number of subjects and scripts in this study are small, and so there is naturally some question about whether this indicates widespread practice. This study was intended to be an in-depth examination of review of marking, albeit across different subjects, different types of item, and all 4 boards operating in England. On this basis, it is likely to have some merit in generalising to other scripts and to other subjects in reviews of marking. The survey is important here as a complement to this in-depth study, including reviewers from all boards, across all subject areas.

6 Survey: introduction and research objectives

In order to understand the extent to which mark changes made during marking and moderation reviews reflect Ofqual's rules, and the extent to which reviewers were trained to undertake reviews, and understand the new rules and how to implement them, 2 surveys were distributed to examine:

- the guidance received by marking and moderation reviewers in how to conduct reviews of marking and moderation;
- the approaches and guiding principles used by reviewers when conducting marking and moderation reviews;
- reviews of marking and moderation mark change judgements in practice.

7 Methodology – review of marking and moderation surveys

7.1 Survey design

The survey for the reviews of marking was based substantially on a survey conducted on Reviews of Marking /Enquiries about Results (Ofqual 2016). The reviews of moderation survey was designed to be analogous to this survey and was piloted with one senior moderator.

The surveys were created and distributed online using SurveyGizmo software. There were 2 surveys, one designed specifically for marking reviewers, and one designed specifically for moderation reviewers. The full surveys can be found in Appendix D and Appendix E, respectively.

Each survey enquired about the guidance (instructions and training) received in conducting reviews of marking or moderation, the guiding principles and approaches used by marking and moderation reviewers to make their judgements, and marking and moderation review behaviours in practice. Each survey required a maximum of approximately 65 responses depending on answers to 'routing questions'.

7.2 Participants

Board examiners of any subject who conducted service 2 reviews of marking or service 3⁸ reviews of moderation in 2016 were invited to complete the survey. In total,

⁸ Service 2 is the terminology exam boards use for a review of marking, where the original marking is reviewed for the whole script. Service 3 is the terminology exam boards use for a review of

there were 1603 marking reviewer, and 255 moderation reviewer responses, with 1295 marking reviewers and 181 moderation reviewers who had conducted reviews of marking in 2016 fully the survey. Approximately a third to a half of all reviewers and moderation reviewers from the 2016 series responded to the survey.

7.3 Respondent descriptives

The marking and moderation reviewers that took part in the survey on average were very experienced in conducting reviews of marking (including when it was formerly known as Enquiries about Results) or moderation.

The marking reviewers that took the survey had been examining on average, for approximately 21 years ($SD = 11$), and conducting Service 2 reviews of marking, on average, for approximately 9 years ($SD = 8$).

The moderation reviewers that took part in the survey had been examining, for approximately 18 years ($SD = 9$), and conducting Service 2 reviews of moderation, for approximately 8 years ($SD = 6$).

The marking and moderation reviewers also had a range of examiner role levels (presented in Figure 15 and Figure 16) which was broadly proportionate to the numbers that hold these roles.

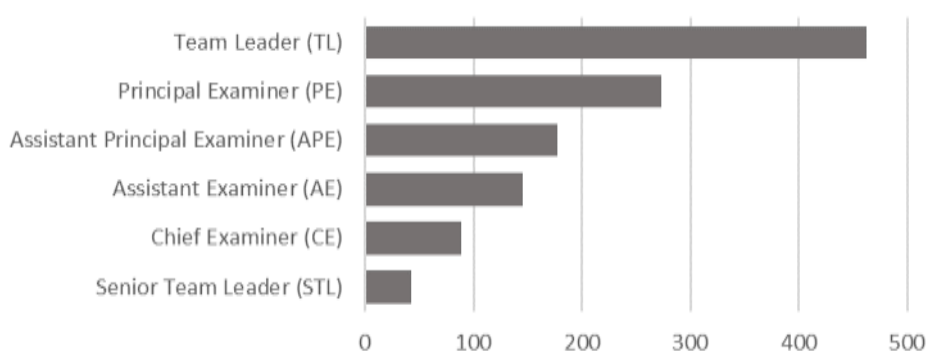


Figure 15. The number of marking reviewers by role.

moderation, where the original moderation is checked for the whole centre. Service 1 is a 'clerical check', generally not conducted by an examiner, to ensure all responses and pages were marked and that all the marks were added up correctly.

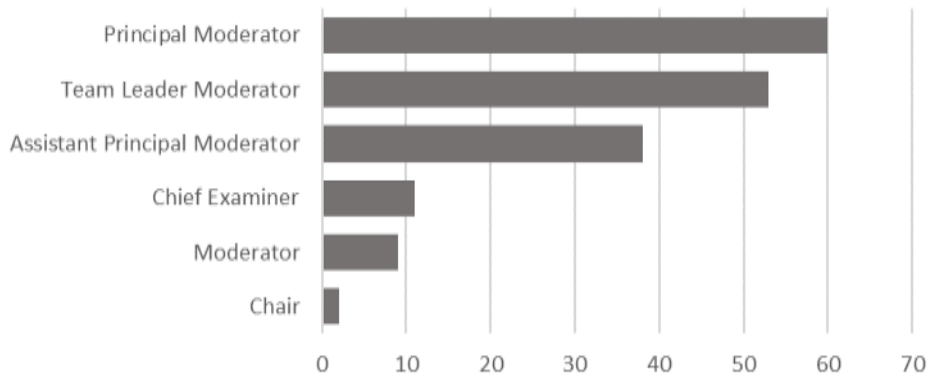


Figure 16. *The number of moderation reviewers by role.*

Marking reviewers indicated that they conducted on average approximately 148 ROMs ($SD = 271$) in 2016, with some reviewers conducting reportedly 3000 reviews in 2016.

Moderators conducted on average 11 moderation reviews ($SD = 11$) in 2016. Because each moderation includes multiple candidates, this smaller number of reviews of moderation is to be expected.

Reviewing markers represented 50 subjects and reviewing moderators represented 29 subjects.

Marking and moderation reviewers undertook reviews on a range of subjects in 2016, the proportions of which are presented in Figure 17 and Figure 18 (also see **Appendix E**). These proportions are broadly consistent with proportions of entries in these subjects. Figure 17 and Figure 18 present the 20 most frequent subjects that marking and moderation reviewers undertook reviews of marking or moderation on in 2016.

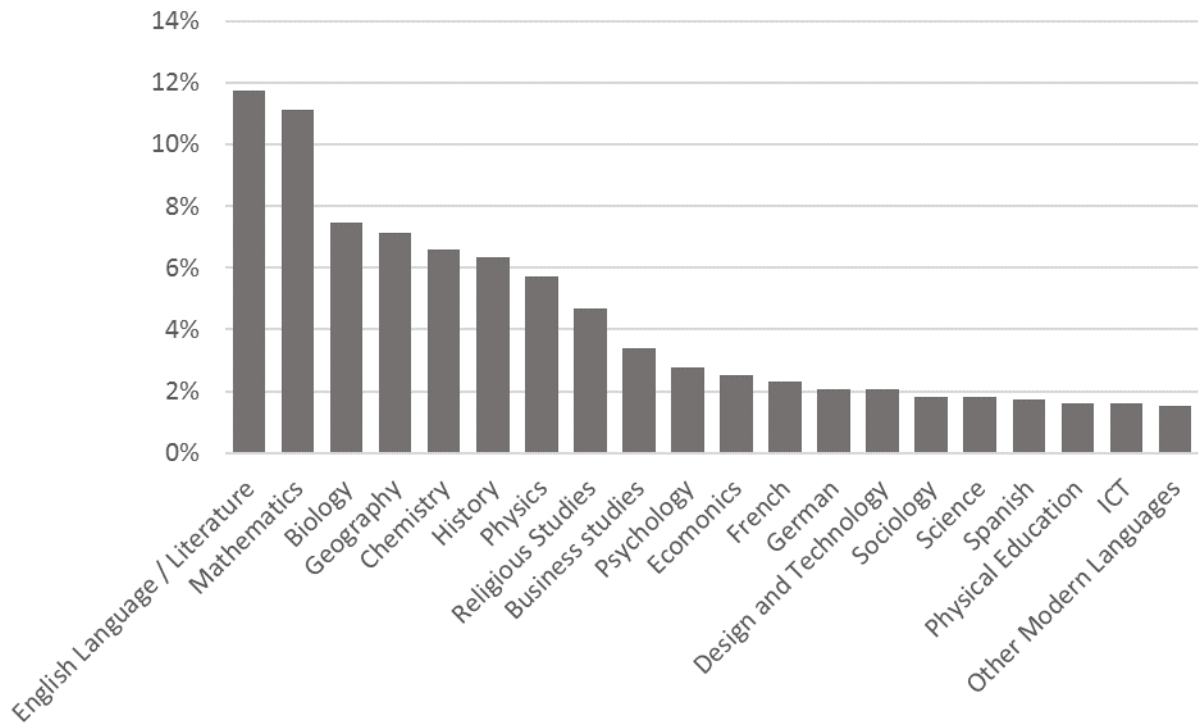


Figure 17. The proportion of reviewers that conducted ROM in 2016 by the 20 most prevalent subjects.

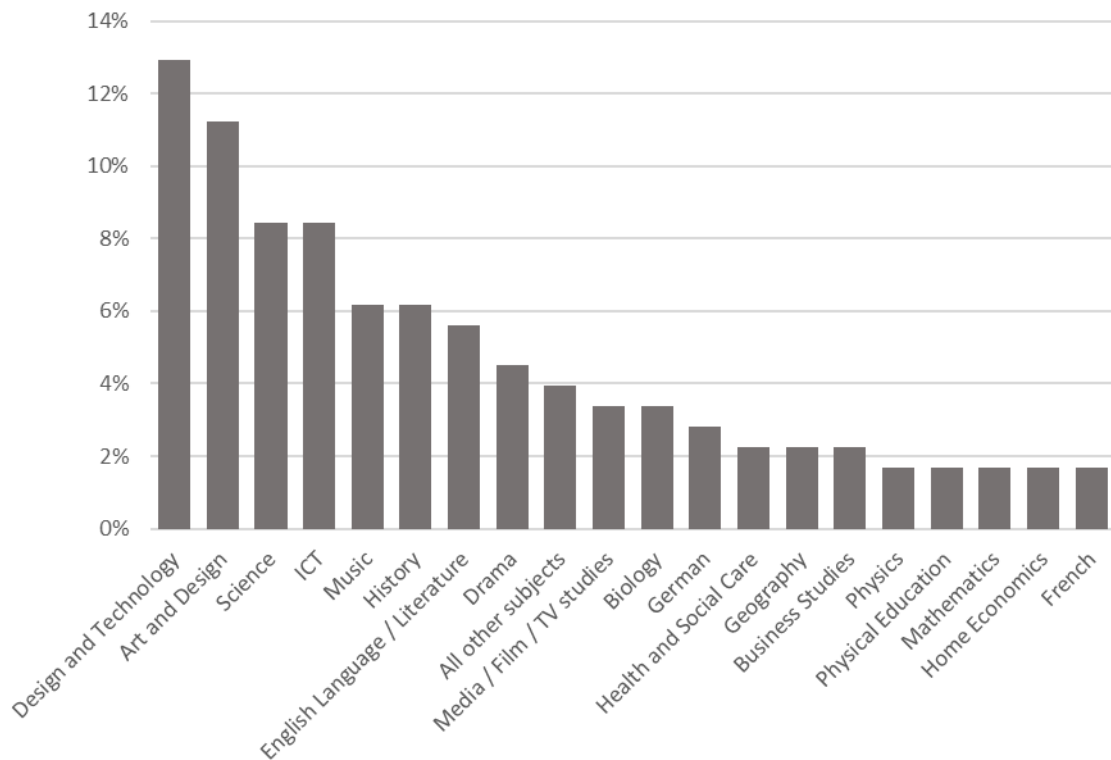


Figure 18. The proportion of moderation reviewers that conducted reviews of moderation in 2016 by the 20 most prevalent subjects.

For reviews of marking, there was an even split of reviewers that conducted ROM at GCSE and AS/A level (52% and 48%, respectively). Whereas, for reviews of moderation, there are more moderation reviewers who conducted reviews of moderation at GCSE than at AS/A level (61% compared to 39%, respectively).

Different types of response are generally marked using different types of mark scheme, namely 'points- or levels-based' mark schemes. Levels-based mark schemes categorise students' responses into marking bands or levels, differentiated by different levels of quality of the work, ability and criteria met. Levels-based mark schemes are typically used to mark longer responses, such as essays. Points-based mark schemes award marks when criteria in the mark scheme have been met in the response. Points-based mark schemes are typically used when there is one, or few, correct responses that meet the mark scheme criteria, and are typically used to mark in objective subjects such as maths and the sciences. The prevalence of points-based and levels-based mark schemes the marking reviewers used was relatively evenly split, 25% and 27% respectively, with 48% of reviewers, indicating that the question papers on which they conduct ROM consisted of both points- and levels-based marking. The moderation reviewers predominantly examined using a levels-based mark scheme compared to a point-based mark scheme, 49% compared to 19% respectively, with 32% of reviewers indicating that the question papers on which they conduct moderation reviews consist of both points- and levels-based marking.

7.4 Fieldwork

The surveys were completed online in December 2016 and January 2017. The surveys were distributed by the 4 exam boards: AQA, OCR, Pearson and WJEC, to their examiners who had conducted either reviews of marking or reviews of moderation, in 2016. Respondents who had conducted reviews of marking *or* reviews of moderation in 2016 were directed to the appropriate survey accordingly. Those who indicated conducting reviews of marking *and* moderation were directed to the reviews of moderation survey. This was because a lower response rate was expected for moderation reviewers. The survey took approximately 15 minutes to complete.

8 Results for the review of marking and moderation surveys

8.1 Reviews of marking survey

8.1.1 Guidance received for reviews of marking

In total, 99% of reviewers indicated receiving instructions in how to conduct ROM and 64% of reviewers indicated receiving training in how to conduct ROM. When asked

how well-prepared they felt and if they fully understood how to conduct ROM, there appeared to be little difference between perceived preparedness of those who had instructions only and those with instructions and training (see Figure 19). The small number (n=7) who said they had had neither, felt less prepared than the other 2 groups.

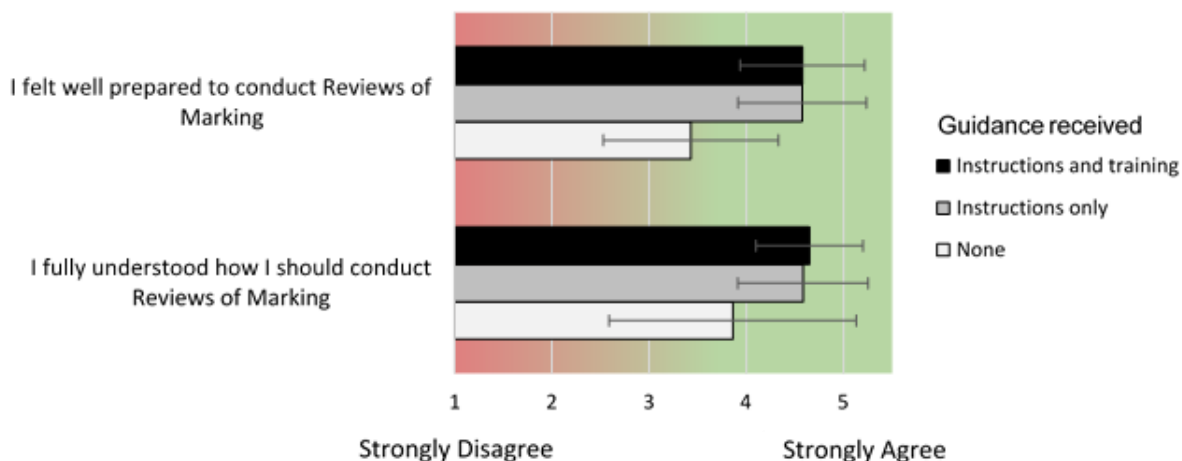


Figure 19: respondents' perceptions of feeling well-prepared and having a full understanding.

Interestingly, when we asked those that had received instructions, whether the 2016 instructions were different from those in the previous year (2015), 89% thought they instructions were different from last year and 11% of marking reviewers indicated that the instructions were the same. This suggests that they were not aware of any changes in the instructions in how to conduct ROM. This indicates that either they had not fully absorbed the instructions in 2016 and/or 2015; or that the instructions' differences were not sufficiently apparent.

Marking reviewers were asked how important it was to read the instructions; to which 74% replied that it was essential.

Regarding the type of training received (Figure 20) the majority of marking reviewers indicated training was done so via online training (42%) and online briefing (18%), and re-standardisation (29%). Any form of face-to-face training or interaction via webinar, for instance, was minimal (7%). The majority of training took 1 to 2 hours to complete.

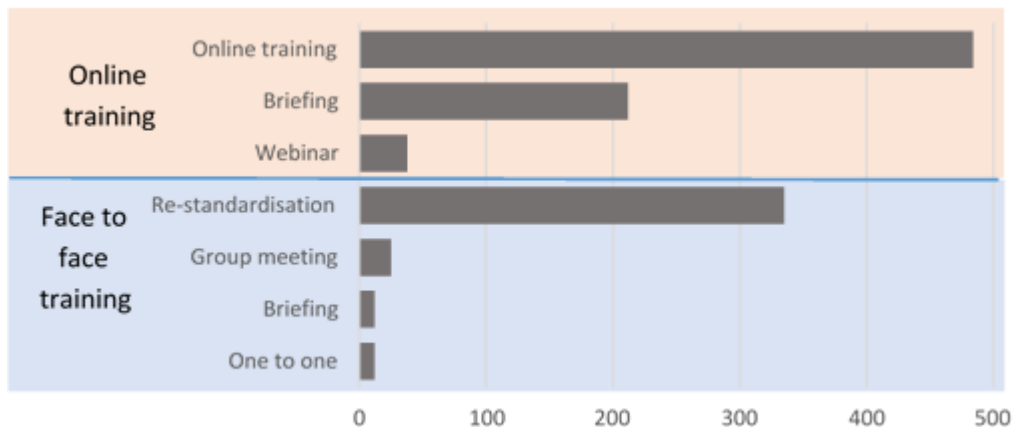


Figure 20. The number of reviewers that received training, by training type.

Note. Respondents could select all training types that they had received.

8.1.2 Understanding the reviewers guiding principles to reviews of marking.

Some survey items addressed how marking reviewers approached RoM, by asking their level of agreement to a series of statements regarding how they understand the RoM rules, and how they would behave. Responses indicated that, in general, there was understanding of the process, however there are some that have misunderstood the aim and purpose of the RoM process, and do not conduct ROM in line with the ROM rules.

There are a number of items in which the respondents should indicate total agreement with the statement. The following statements are such examples where less than 100% of respondents strongly agreed:

- 'In the Reviews of Marking process, I review each response carefully to make sure the original examiner has not missed anything' (80% strongly agree);
- 'When conducting Reviews of Marking, I view the process as a review of the original marking: if the original mark could be justified I do not change it' (75% strongly agreed).



Figure 21. The distribution of responses indicating agreement with statements relating to understanding of the Review of Marking process.

Note. Shaded regions indicate the 'correct' answer, where appropriate.

Similarly, there are some survey items in which the respondents should indicate strong disagreement. Examples (strong disagreement figures in brackets) include:

- when conducting Reviews of Marking, I believe the mark I give should be different from the original mark; (83% strongly disagreed);
- in Reviews of Marking, when the original mark is justified I believe that my mark should still override the original mark; (67% strongly disagreed); and
- in Reviews of Marking, I try to find a few marks for the candidate.(78% strongly disagreed)

In each of these cases, the majority of respondents supplied the best possible answer, but the distribution of responses for these items indicates that there are a number of reviewers who do not understand and are unlikely to be conducting reviews of marking in line with Ofqual's rules.

8.1.3 Making mark adjustments in reviews of marking

Although marking respondents had previously indicated they generally understood the RoM process, we wanted to examine how this was reflected in their reported behaviour in conducting RoM in practice. Respondents were presented with a series of scenarios, to which marking reviewers indicated whether they agreed or not with making adjustments to the original examiner's mark.

Encountering different sources of marking error

Several survey items addressed how marking reviewers approached RoM. Reviewers were given examples of circumstances they may experience during reviews of marking and asked if they would change the mark in these circumstances. First, reviewers were asked how frequently they had encountered each scenario (see Figure 22) and then (only if they had indicated if they encountered it) they what they do about it (see Figure 23).

Figure 22 shows that, on average, no errors were deemed to appear 'very often' and mostly they appeared sometimes or occasionally. The most frequent error was the original examiner missing a creditworthy point.

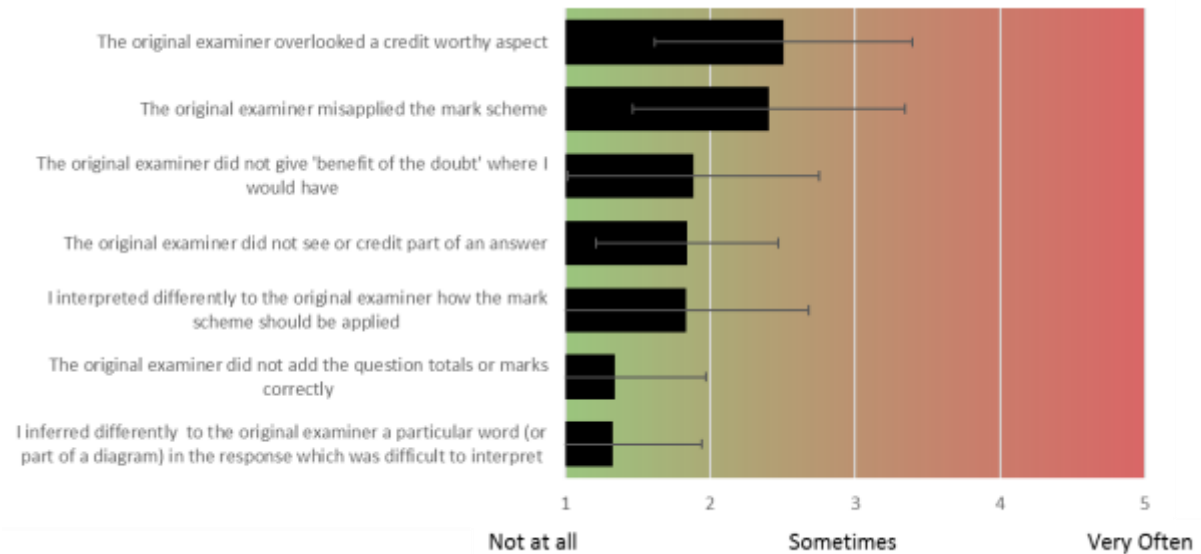


Figure 22: Mean frequency of encountering particular marking errors when conducting reviews of marking. Bars show the standard deviation of responses.

Figure 23 shows the distribution of responses to what reviewers would do in the different marking error scenarios and whether they would change the mark. A high percentage (85%) of reviewers correctly indicated they would definitely change the mark where marks were not added correctly, but not 100%. In other scenarios which should be fairly clear cut in changing the mark (misapplication of mark scheme, not seeing/crediting part of answer), there is less agreement. This indicates that in a small number of cases reviewers are not clear on when and when not to change marks.

The key scenario which respondents gave an unexpected answer to, and one which does not adhere to Ofqual’s rules is to do with ‘benefit of doubt’. Benefit of doubt refers to deliberately giving a slightly more positive interpretation to what is written, and awarding accordingly. A large proportion (50%) of respondents indicated that they would change the mark to give a candidate ‘benefit of doubt’, even though this means they would be replacing one legitimate mark with another legitimate mark. This surprising result, that respondents admitted to it, is supported in some part by the findings of the Marking Review Study in Section 3 of this report.



Figure 23. The distribution of responses as to how likely it would be for the marking reviewers to change the mark in each circumstance.

Note. Marking reviewers were presented with each circumstance if previously reported that they came across it. Shaded regions indicate the 'correct' answer.

Points-based marking scenarios

Reviewers who indicated that they conducted reviews of marking on papers assessed with points-based mark schemes broadly indicated that they would conduct RoM in line with the RoM rules. The distribution of responses for each scenario are presented in Figure 24. The majority of reviewers agreed they would change marks when creditworthy responses had been missed, and when responses had been credited when they should not have. Interestingly, there was lesser agreement to leave the original marks unchanged in scenarios in which the original marker and the reviewer were both correct in their interpretation to award 2 different marks. This indicates they would replace one legitimate mark with another legitimate mark.

Again, although the picture on the whole is positive, we observe that there are a number of reviewers in all 4 scenarios who responded in a manner that goes against the RoM guidelines. For instance, some reviewers disagreed to changing the mark where the original marker credited a response when they should not have, or missed a credit worthy aspect of the response.

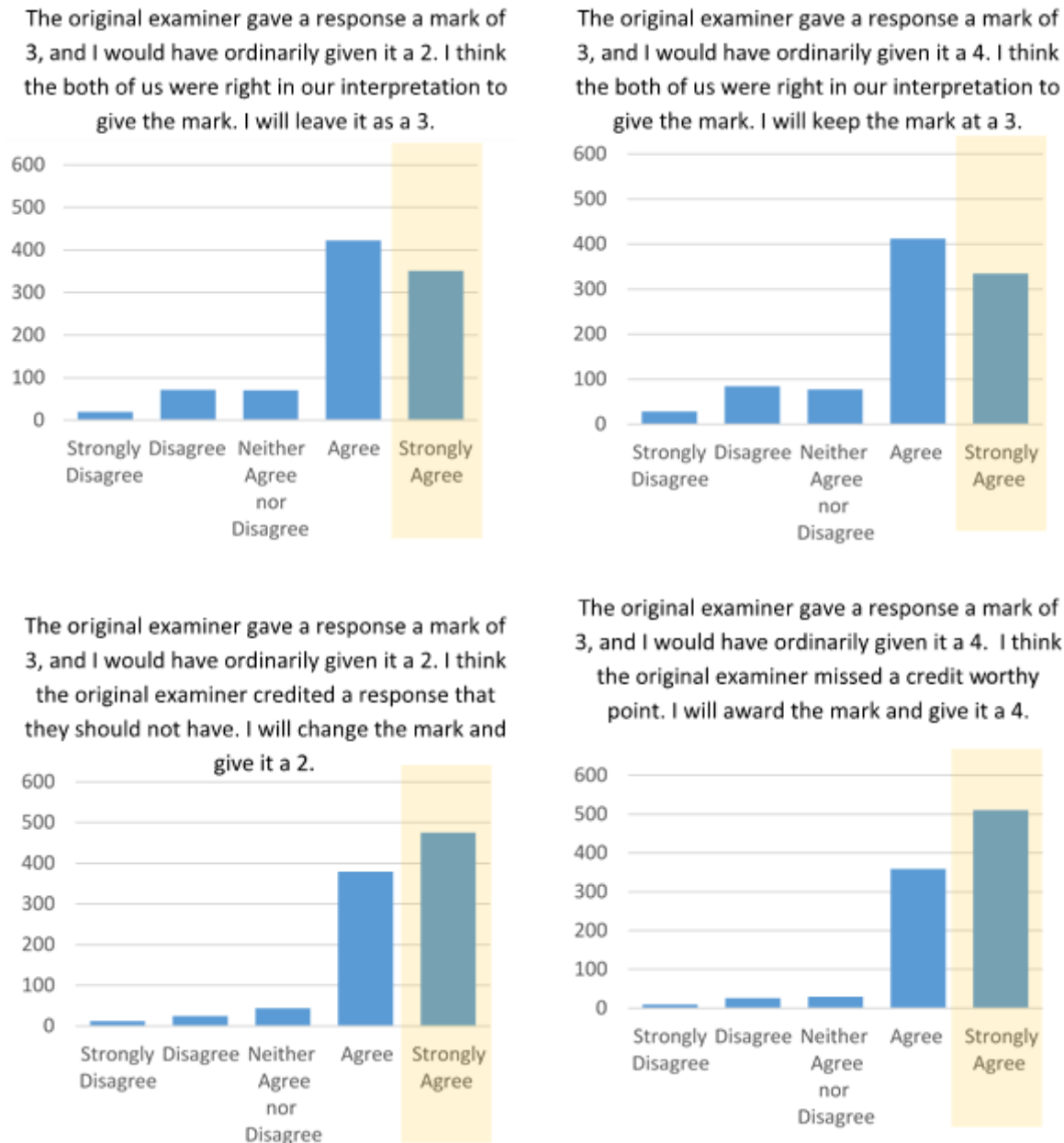


Figure 24. The distribution of responses as to how much examiners agree to each points-based marking scenario.

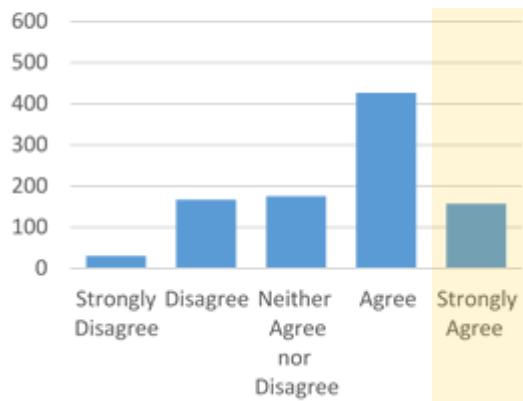
Note. Shaded regions indicate the 'correct' answer. Marking reviewers were presented with the points- and/or levels-based scenarios if they previously indicated that they conducted reviews of marking on scripts comprising of these types of mark schemes.

Levels-based marking

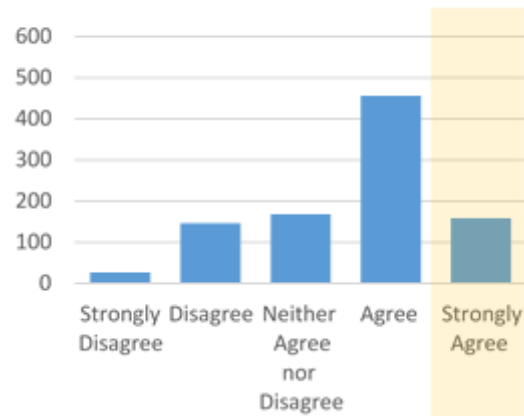
Reviewers disagreed with leaving marks unchanged if the mark they would have awarded was in a different level/band, irrespective of whether the mark increased or

decreased. The spread of responses (Figure 25) indicate that the majority of reviewers would change the mark if it was to a different band. When the mark that the reviewer would have awarded was different but in the same band/level to the original mark, the responses were more spread in terms of whether reviewers agreed to leave the mark or not.

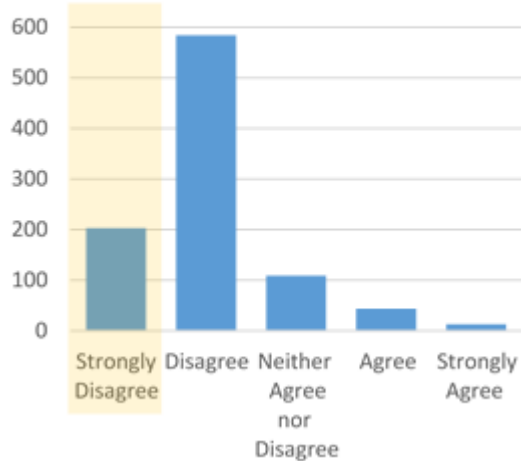
The original examiner gave a response a mark of X. I would have ordinarily given a **higher** mark but in the **same level/band**. I will leave the mark as X.



The original examiner gave a response a mark of Y. I would have ordinarily given a **lower** mark but in the **same level/band**. I will leave the mark as Y.



The original examiner gave a response a mark of W. I would have ordinarily given a **higher** mark but in a **different level/band**. I will leave the mark as W.



The original examiner gave a response a mark of Z. I would have ordinarily given a **lower** mark but in a **different level/band**. I will leave the mark as Z.

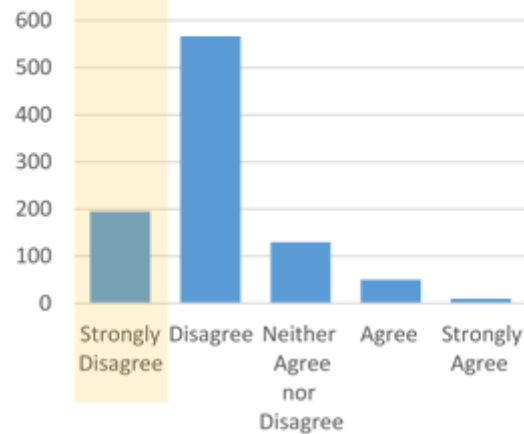


Figure 25. The distribution of responses as to how much examiners agree to each levels-based marking scenario.

Note. Shaded regions indicate the 'correct' answer.

Marking reviewers further indicated that they generally spend a bit more time on RoM than on the original marking; they believe the quality of their marking is the same or a bit better than the original examiner's, and; they find it easy, or neither easy nor difficult to understand how the original examiner justified their marks (see Figure 26, Figure 27 and Figure 28).

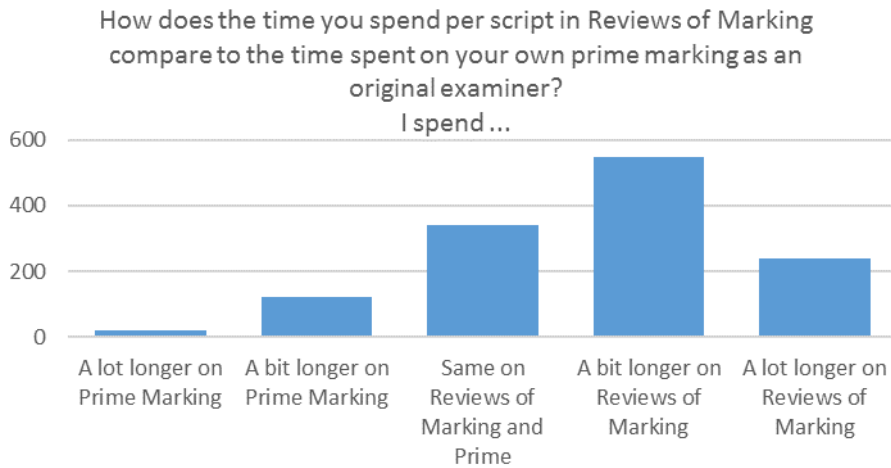


Figure 26. Responses indicating the time reviewers spent on ROM.

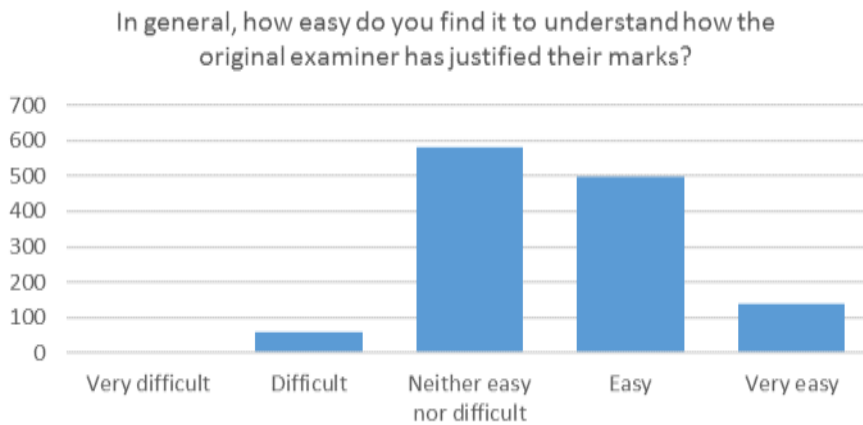


Figure 27. How marking reviewers' compare their own marking with the original marking.

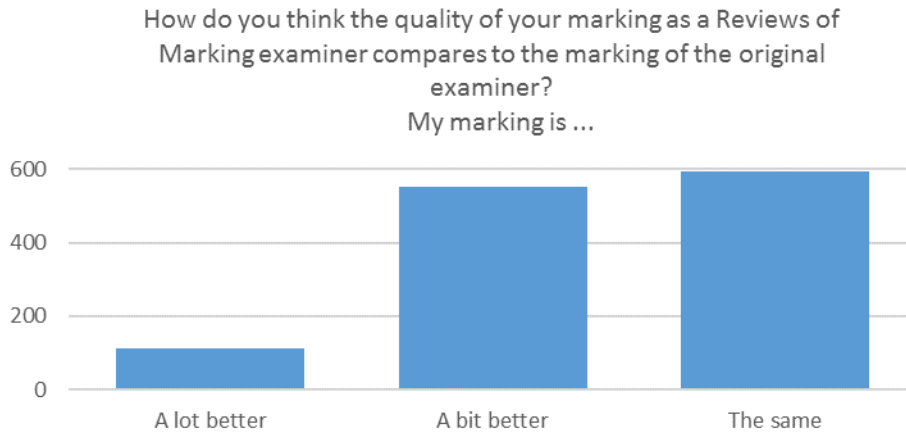


Figure 28. How easy marking reviewer find it to understand how the original mark was justified.

8.2 Reviews of moderation survey

8.2.1 Guidance received for reviews of moderation

While 99% of moderation reviewers reported receiving instructions in how to conduct reviews of moderation, only 46% reported receiving training. Similar to the reviews of marking survey, those who had received the training did not report feeling better prepared than those with only instructions (see Figure 29).

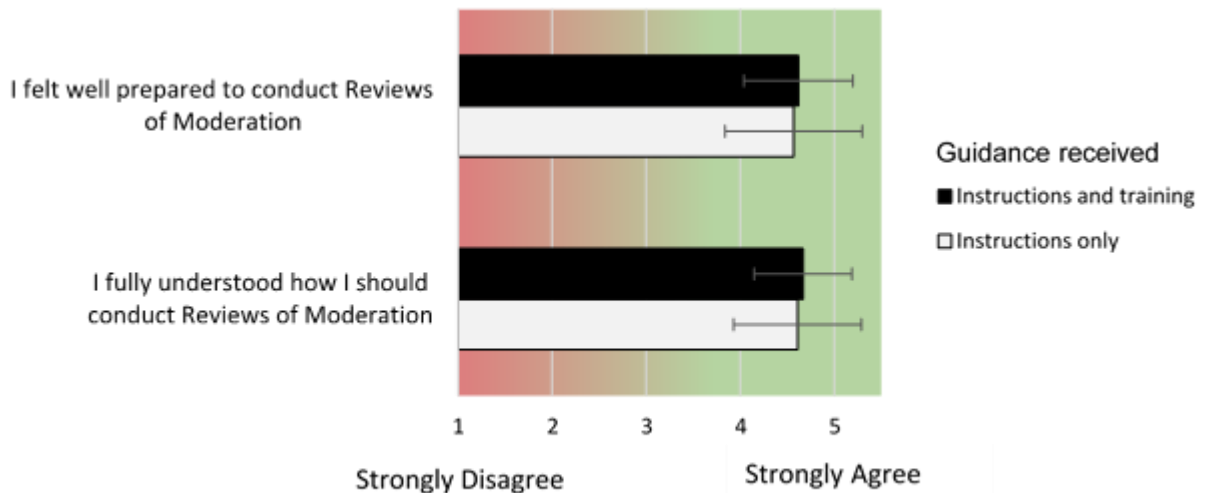


Figure 29. Average scores for agreement in feeling prepared and understanding how to conduct reviews of moderation, by whether instructions and/or training was received.

For those that had received instructions, when asked how different the instructions were from the previous year (2015), 82% of moderation reviewers indicated that the instructions were different from last year, while 18% of moderation reviewers indicated that the instructions were the same. In other words, they were not aware of any changes in the instructions in how to conduct reviews of moderation.

Moderation reviewers reported the type of training they received (Figure 30). The majority of moderators who experienced training did so via an online medium. Forms of interactive training such as face-to-face training or interaction via webinar, for instance, were reported by just over a quarter of the moderators who had received training (29%).

The majority of training took less than an hour (31%) or 1 to 2 hours (36%) to complete.

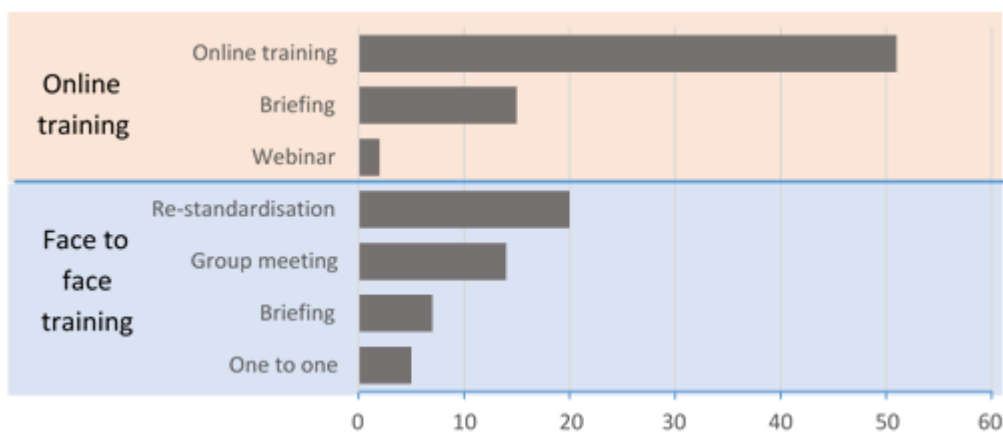


Figure 30. The number of moderation reviewers by the type of training received.

Note. Reviewers selected all training types that they had received.

8.2.2 Understanding the approach to reviews of moderation

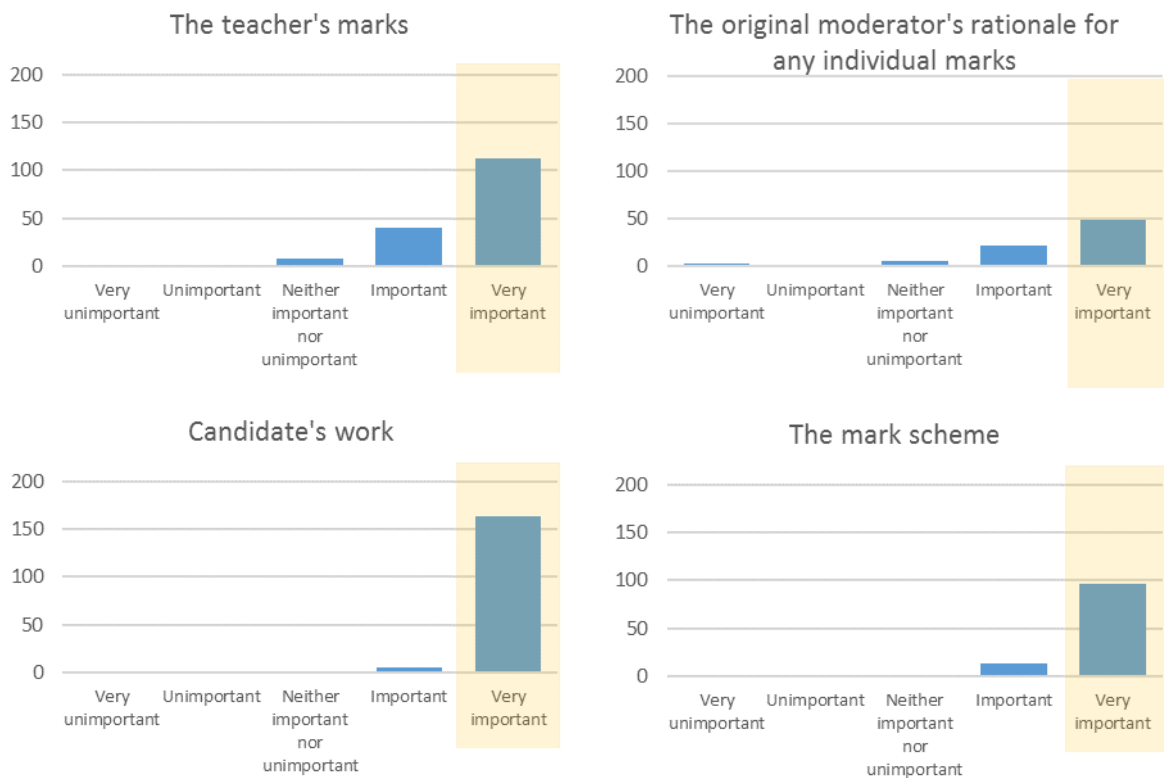
Moderation reviewers were asked how important some sources of information were for influencing their judgements in the marks awarded. The distribution of responses as to the importance of these sources of information are presented in Figure 31.

There appears to be some very different practices in terms of what sources of information are considered important in conducting a review of moderation. This indicates that there are different methods being undertaken by moderation reviewers, with different sources of information, some that should not be taken into consideration (centre name, centre number and candidate name) influencing the judgements in awarding marks during reviews of moderation. These findings

corroborate those reported in the research on moderation published by Ofqual (Ofqual, 2017b)⁹

There are some interesting points to reflect on here:

1. Some moderators (around 40%) indicated that the original moderator's rationale for individual marks was very important or important for making moderation decisions. If review of moderation is deciding upon whether the original moderator marks are legitimate, this seems to be an important source of information.
2. A similar percentage (around 40%) thought that knowing the original adjustment to the centre was important or very important. Again, if deciding whether the original moderation produced a legitimate outcome for a centre, this would seem to be a vital source of information, though it is possible that another part of the process employed by boards checks the adjustment.



⁹ <https://www.gov.uk/government/publications/exam-and-assessment-marking-research>

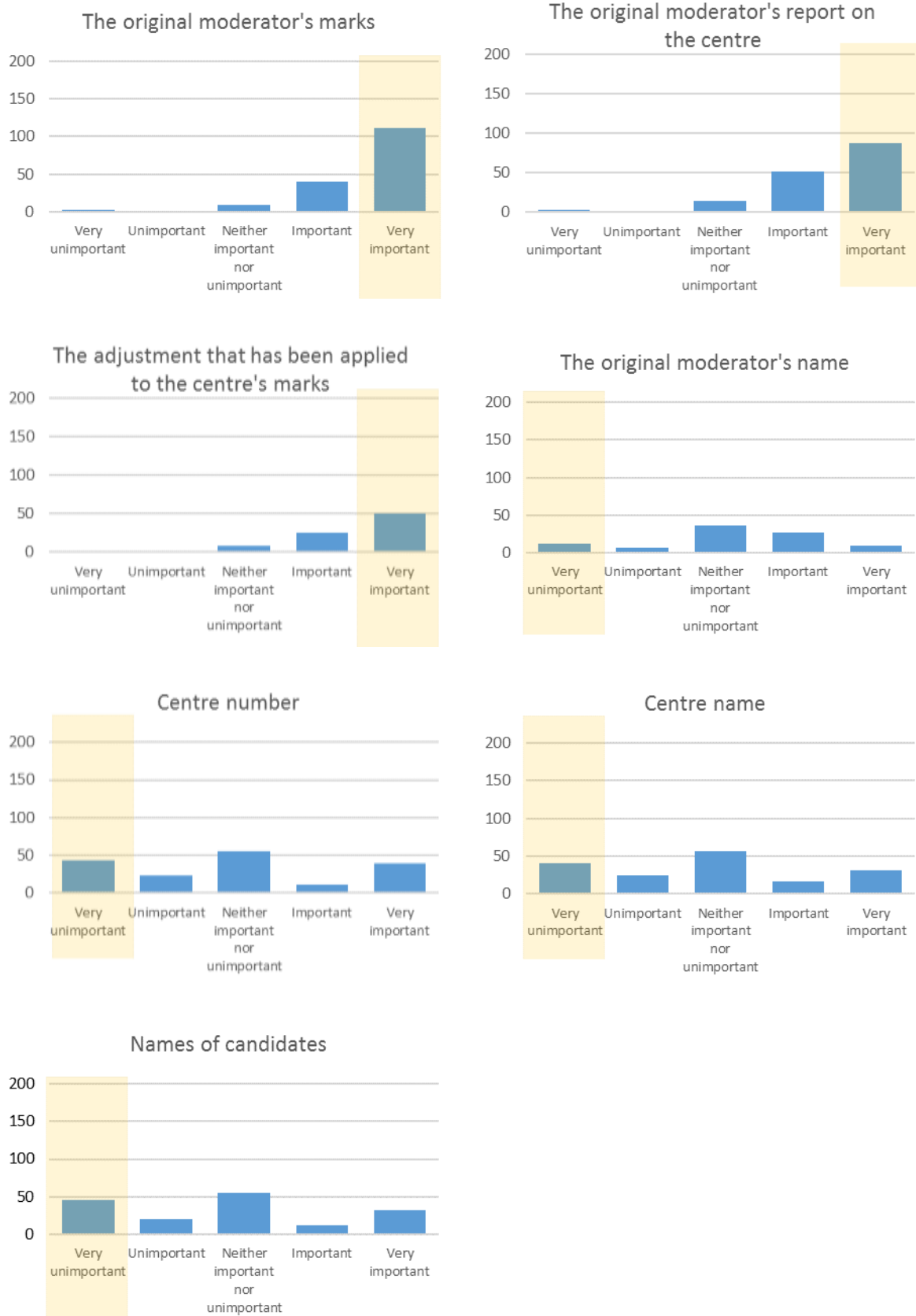


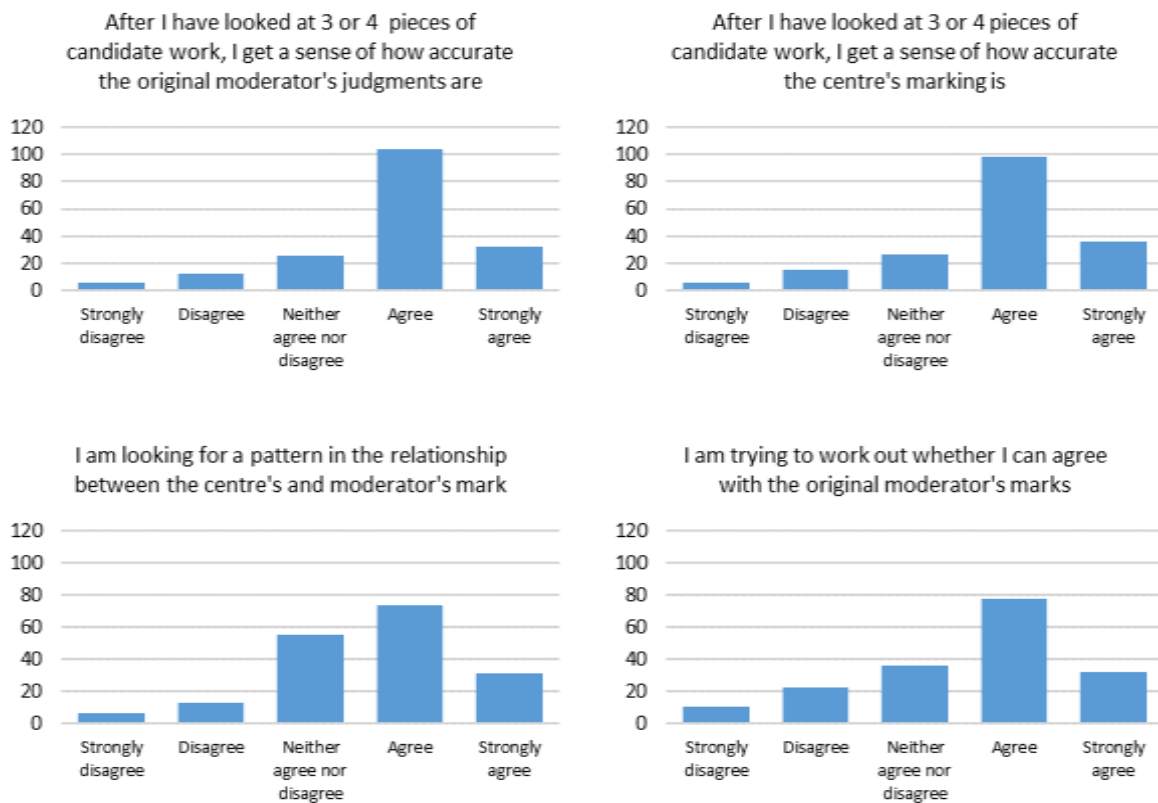
Figure 31. The distribution of responses as to how important certain sources of information are in influencing judgements in reviews of moderation.

Note. Respondents could indicate their perceived importance of sources only if previously indicated that they have this information available. Shaded regions indicated the 'correct' answer.

8.2.3 Making mark adjustments in practice in reviews of moderation

To further understand how reviews of moderation are approached, moderation reviewers were asked their agreement to a series of statements (see Figure 32) in how they use the teacher's and moderator's marks, and the candidate work to make judgements on making mark changes.

Moderation reviewers show a range of responses to these questions (see Figure 32 for details of each statement), again indicating different approaches are being used by moderation reviewers when making judgements.



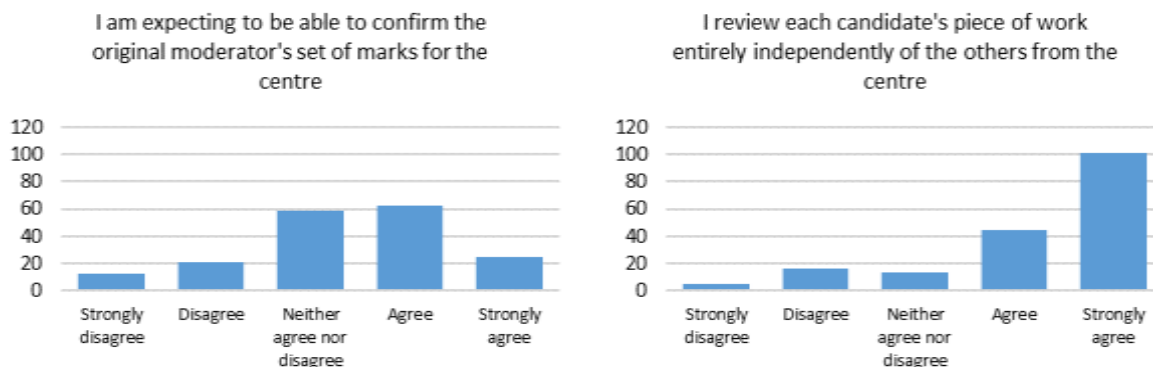


Figure 32. *Distribution of responses regarding how moderation reviewers use the teachers' and moderators' marks and the candidate work to make judgements on making mark changes.*

8.2.4 The role of tolerance in reviews of moderation

When moderating teacher's marks, if the moderator's mark adjustments are out of tolerance¹⁰, the teacher's marks would be subject to mathematically calculated mark adjustments. Therefore, if the teacher's marks are out of tolerance, the moderator is liable to make changes. Moderation reviewers were asked to report their agreement to a series of statements that examined how their judgements in making mark changes are influenced by the tolerance (Figure 33).

The majority of moderation reviewers correctly indicated that they would give a mark out of tolerance of the original moderator's marks if the candidate's response warranted it. For other items, there is more of a mixed picture in how tolerance influences mark judgements in a moderation review and may indicate that there are different approaches being taken towards awarding marks out of tolerance. This is particularly noticeable when the original moderation is generally fine, whereby some moderation reviewers would correctly give a mark out of tolerance, and some would not, in error. This is a perennial dilemma for moderators, whether they are making individual judgements about individual pieces of candidate work; or whether the marks should be mediated by a more holistic take on the quality of the marking in the centre, and of the original moderator.

¹⁰ Tolerance is effectively an 'allowed' mark difference – a teacher mark within tolerance or the moderator mark is unlikely to trigger a centre adjustment. At a review of moderation, if the reviewing moderator marks are within tolerance of the original moderator's marks, then the original moderation decision is likely to stand.

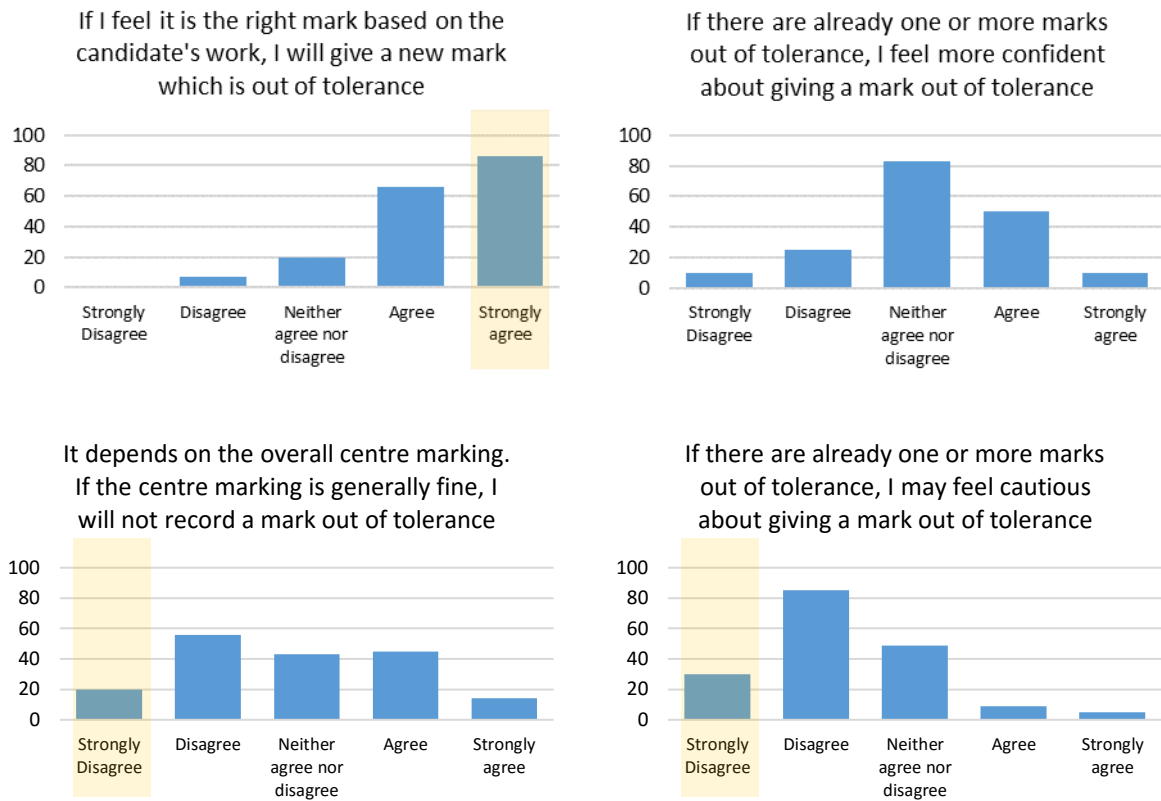


Figure 33. The distribution of responses to agreement to statements in how to deal with tolerance in reviews of moderation.

Note. Shading indicates the 'correct' answer.

How Moderation reviewers' make judgements in which mark to record when faced with original moderator mark changes that are in and out of tolerance was also assessed (see Figure 34).

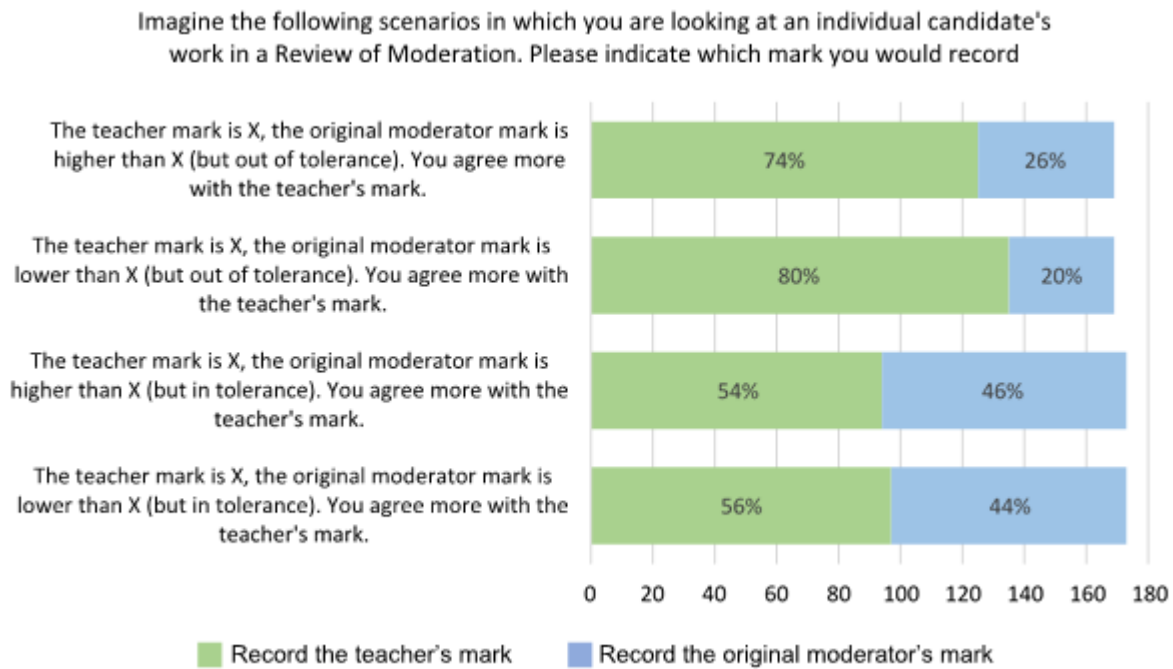


Figure 34: *which mark to record – original moderator or original teacher mark – in different scenarios*

The moderation reviewer was more likely to correctly record the teacher's mark where they agreed with that mark, regardless of whether the mark would increase or decrease at the moderation review. The responses to these scenarios however, do cause some concern. Between a fifth and quarter of moderation reviewers would record the moderator's mark, despite agreeing with the original teacher mark.

9 Findings and conclusions – review of marking and moderation surveys

The aim of the surveys was to aid in the understanding of:

- (i) the extent to which mark changes made during marking reviews reflect Ofqual's rules, and;
- (ii) the extent to which reviewers were trained to undertake reviews and understand the new rules and how to implement them.

Marking and moderation reviewers' responses appeared to show, in general, good understanding of Ofqual's rules and guidelines in how to conduct reviews of marking and moderation. Interestingly, and in some instances causes for concern, marking and moderation reviewers reported using a range of guiding principles and approaches when conducting reviews of marking and moderation. When responding to scenarios in practice, there was a lesser consensus as to whether mark

adjustments should be made or not, with many marking and moderation reviewers making mark changes in a manner that goes against the RoM guidelines. It appears therefore, that there may have been discrepancies between the reviewers' understanding of the RoM guidelines, and how this is implemented in their reviewing behaviour. For instance, despite marking reviewers reporting that they were most likely to leave marks unchanged if the original mark was justified, and were most likely to not try and find a few additional marks for the candidate, 50% of marking reviewers nevertheless reported applying 'benefit of doubt' where the original examiner did not. Between a quarter and fifth of moderation reviewers also reported that despite agreeing with the teacher's mark, they would leave a moderators mark that was out of tolerance.

We have to make reasonable assumptions that the responses to the survey questions reported here provide an accurate representation of the thoughts and behaviours of the exam board's examiners who took part, offering credible implications to be drawn. However, we do note that despite efforts in clarity of the survey questions it is possible that some reviewers may have mis-read or mis-interpreted aspects of the questions and response options. It is also possible that respondents provided responses which were not an entirely true representation of their reviewing behaviour because of socially desirability or the passage of time between the survey and the reviewing.

With regards to the training and guidance received in how to conduct reviews of marking and moderation, although 99% of reviewers reported receiving instructions in how to conduct reviews of marking or moderation, only 46% of moderation reviewers and 64% of marking reviewers said they had received any training prior to undertaking reviews. Reviewers further indicated that the training they received was predominately conducted online through online training and online briefings. Although re-standardisation was the most likely form of face-to-face training, the prevalence of more intimate and interactive training in the form of webinars, and group- or one-to-one meetings was minimal. Marking and moderation reviewers indicated that the training took, for the majority, 1 to 2 hours to complete.

Marking or moderation and reviewers' generally agreed that they felt prepared to conduct review of marking and understood how they should conduct reviews of marking or moderation. However, those that had received only instructions in how to conduct reviews of marking and moderation reported understanding how to conduct reviews of moderation and being well-prepared as much as those who received both instructions *and* training. It therefore appears that the training did not improve reviewers' feelings of preparedness and understanding in the marking and moderation review process beyond that of the instruction documentation.

For some, the implementation of the new guidelines in conducting reviews of marking and moderation were reported here as disagreeing with Ofqual's guidelines. This may, at least in part, be due to the instruction documentation not highlighting clearly

the differences between the old and new guidelines with one-tenth of marking reviewers and one-fifth of moderation reviewers reporting that they believed the instructions they received in 2016 to be no different than in 2015. Moreover, only three-quarters of reviewers believed that it was essential to read the marking and moderation review instruction documentation prior to undertaking any reviews, meaning that despite the new guidelines being available on the instruction documentation, a number of reviewers may not have read them.

10 Overall conclusions and discussion

Our original questions were as follows:

1. The extent to which mark changes made during reviews of marking reflect Ofqual's rules, ie to only correct error.
2. The extent to which reviewers were trained to undertake reviews and understand the new rules and how to implement them.

This research, drawing upon both the review of marking study and the survey of reviewers indicates the following answers:

- The vast majority of items in the study (93%) were reviewed according to Ofqual's rules - such that marking error and only marking error was corrected. But a proportion of items (6%) were not. This impacted overall upon 60% of scripts in the study¹¹. There were noticeable subject differences – with 45% of maths scripts, 60% of English scripts and 80% of biology scripts affected. While most of the discrepancies between the live review of marking and the definitive review of marking were very small, not all of them were;
- in part, this is answered by the outcomes provided in question 1 – there were reviews that did not reflect an implementation of the new rules. This small scale in-depth study of 3 subjects is complemented by the larger scale survey which found that while the majority of examiners (64%) received guidance and training on conducting reviews of marking, and reported that they understood the new rules, when given more specific scenarios of whether or not they would change the marks, some gave responses indicating that they would not implement those principles. There is some disjunction between understanding the high level principles and understanding under which scenarios to implement them. For example, 50% of those in the survey said they would 'definitely' change the mark to give 'benefit of doubt'. In reviews of moderation,

¹¹ This is because even if 19 of the 20 items in the paper were marked/reviewed correctly, a mark change to just one item will change the script mark.

only 46% of moderators indicated they had received training. Moderator responses were often more divided on what to do in certain scenarios indicating that perhaps their task (reviewing the original moderation) is not entirely clear to them.

Looking at the review of marking and moderation data¹² there is an indication that the amount of grade change and small mark change is less in 2016 than in 2015, indicating that exam boards have implemented the new rules. This report also supports the view that boards have implemented the new rules, but that the implementation in this first year has not yet been fully realised. It may be that exam boards and reviewers and moderators will continue to improve in this regard. One issue is around training, where 36% of reviewers and 54% of moderators said they had not received training (or did not perceive that they had been trained). Those that had received training did not feel any more prepared than those who had received instructions only. It seems likely that both the provision of training and the nature of the training could be improved. Where training contains both principles, and specifics of how implement (eg by mark scheme type, by subject, by example), it is likely to be more successful in affecting behaviour.

The subject differences in the study raise some interesting issues. For mathematics and biology, identifying the 'definitive RoM mark' was possible in very nearly all instances (there were no mathematics items and only 2 biology items (equivalent to 0.1%) where the study could not identify a definitive RoM). Ultimately, this means that the identification of marking error and correction of is a viable and realistic course of action. The rules can be implemented.

But what about English literature? Of the items in the study, it was not possible to determine a definitive RoM mark for around 11% of items. In some cases, (6% of all English literature items), the reviewers and experts could not even decide on whether or not to increase or decrease the mark compared to the original marking. Does this imply a different approach for these more subjectively marked subjects where there is greater 'definitional uncertainty' (Black and Newton, 2016), in other words, there is some imprecision in the definition of what is being measured. One particular issue observed in English literature responses, typically essays, in this 'hard to mark' category was how markers should deal with responses which contained passages which were wrong or irrelevant, whether to ignore or whether to incorporate into the overall judgement could make a big mark difference. What must be remembered is that this is an issue for prime marking, and issues in prime marking in terms of

¹² <https://www.gov.uk/government/publications/exam-and-assessment-marking-research>

unacceptable definitional uncertainty, cannot be fixed at review. It needs to be resolved earlier, at mark scheme construction, at standardisation etc.

Having said that, the mark scheme had sufficient definitional certainty for 89% of English literature items and it was possible to come to a definitive RoM mark. And in the meetings it was observed that the 2 independent pairs readily agreed.

To conclude: The analysis of overall RoMM data shows that the new rules have had an impact on the number of small mark changes this year. The findings of this research shows that examiners reviewed more than 90% of responses in the research consistently with the new rules, such that only marking errors were corrected. However, the study also found that in the sample in the study 6% of items were not reviewed consistently with the new rules, which created potential discrepancies in the review decisions of 60% of students' scripts in the study. The potential for such discrepancies to occur is also reflected in survey responses from a sample of last year's reviewers. While many reviewers gave answers consistent with properly applying new rules, some reviewers did not, especially in relation to giving 'benefit of doubt'. While boards had provided training (required by Ofqual rules), it seems that in the first year of new rules the implementation has occurred but has been partial. This highlights the need for effective training (as well as monitoring) in conducting reviews, and there is more for boards to do in this regard. Finally, the research also indicates that it is possible to apply the new rules and to be definitive about marking error and the correction of it in the vast majority of cases. Where this was more challenging (in English literature) this was more an issue stemming from the interaction between the mark scheme and the nature of the response.

References

- Ahmed, A. and Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy and Practice*, 18:3, 259-278.
- Black, B & Newton, P. (2016). *Tolerating difference of opinion*. Paper presented at the 17th Annual AEA-Europe Conference, Cyprus.
- Marton, F. and Saljo, R. (1976). On qualitative differences in learning: 1 – Outcome and process. *British Journal of Educational Psychology*, 46, 4-11.
- Ofqual (2015). *Research into alternative marking review processes for exams*. Coventry, England. Accessed at: <https://www.gov.uk/government/publications/alternative-marking-review-processes-for-exams>
- Ofqual (2016) *Reviews of marking and moderation for GCSE and GCE: summer 2016 exam series*. Coventry, England. Accessed at: <https://www.gov.uk/government/statistics/reviews-of-marking-and-moderation-for-gcse-and-a-level-summer-2016-exam-series>
- Ofqual (2017a). *Reviews of marking and moderation: Subject level analyses. Summer 2016 exam series*. Coventry, England. Accessed at: <https://www.gov.uk/government/publications/exam-and-assessment-marking-research>
- Ofqual (2017b). *An exploratory investigation into how moderators of non-examined assessments make their judgements*. Coventry, England. Accessed at: <https://www.gov.uk/government/publications/exam-and-assessment-marking-research>
- Suto, W.M.I. and Greatorex, J. (2008) *What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process*. *British Educational Research Journal*, 34:2, 213-233.

Appendices

Appendix A - Overview and Overall Instructions for Subject Experts for the home-based task.

Introduction – what is this study all about?

Ofqual is conducting a general evaluation of Reviews of Marking – service 2 (formerly known as Enquiries About Results). We are conducting this research in order to understand the different methods used in service 2 Post Results Enquiries that have been submitted for a results enquiry in 2016. The findings from this study, along with other impact analysis, will feed into the general overview of Reviews of Marking evaluation, in order to inform of any additional guidance or other changes that need to be implemented in the future.

Ofqual is working collaboratively with 4 boards for this study and 3 question papers/units. This is very useful because it means each board is engaged with this evaluation, as well as giving us insight as to how the Reviews of Marking process might work on questions papers with different styles and subject matter.

So, really, many thanks for agreeing to take part in this study, the results from this study will be a very important part of helping us evaluate and implement changes to the Reviews of Marking process.

Who is taking part in the study?

Taking part in this study session are 2 examiners who have undertaken Reviews of Marking on this unit in the 2016 session, and 2 Subject Experts (including yourself), who have not previously marked on this unit. All 4 participants will be looking at a small number of scripts for this study.

What is my role in this project as a subject expert?

We know you may not have marked on this panel or conducted reviews of marking before. However, we hope that your subject expertise will help us understand the extent to which marking and reviewing decisions are cut and dried and which are more finely judged decisions.

Once you have familiarised yourself with the assessment materials (question papers and mark scheme) for this particular unit, the main task is to look at each of the 10 scripts. In turn, for each item on each script, given all the information to hand, ask yourself:

“are there other correct or plausible marks for this item? Or is this the only possible correct or plausible mark?”

NB You may think that the mark awarded is wrong, in which case, note this and what mark or marks would have been correct or plausible.

We have provided a 'Plausible Marks' excel spreadsheet with which to record your responses to the above key questions. This spreadsheet will ask you to identify any other marks that may be plausible given all of the information at hand, and reasons/explanations for this plausibility. Please see the 'Instructions and Example' worksheet before you start this task.

Subject unit

This study uses [XXX] unit from summer 2016. Please take time to familiarise yourself with the question paper, mark scheme and standardisation scripts before you start the Plausible Marks task.

After you have completed the Plausible Marks task please familiarise yourself with the Review of Marking instructions document.

Scripts

The scripts selected for this study were all scripts that went through the Review of Marking process in 2016 and are broadly representative of all those scripts that went through this process for this unit. You are provided with 10 scripts on which to carry out your Plausible Marks task prior to the meeting day.

OVERALL INSTRUCTIONS

Reviews of Marking – home-based task

1. This work is **confidential**. Please ensure you keep all the scripts confidential as well as all other materials associated with this study. Also, please ensure that you do not discuss this study with anyone other than those directly involved in the study.
2. Use the step-by-step ticklist as you work through.
3. Familiarise yourself with the question paper, mark scheme and standardisation scripts.
4. Complete the Range of Plausible Marks task.
5. Familiarise yourself with the Review of Marking Instructions.
6. Please do not discuss your scripts with other examiners. They may or may not have the same scripts as you.

STEP BY STEP TICK LIST		
	Home-based task: Plausible Marks	✓
1.	Familiarisation with the question paper and mark scheme	<input type="checkbox"/>
2.	Familiarisation with the standardisation scripts	<input type="checkbox"/>
3.	Complete the Plausible Marks task and complete the excel sheet called 'Plausible Marks'.	<input type="checkbox"/>
4.	Send the filled in 'Plausible Marks' excel file via email to [contact] by [date]	<input type="checkbox"/>
5.	Return all scripts back into original pack envelope and bring them with you to the unit's meeting day in Coventry (this one is very important ☺).	<input type="checkbox"/>
6.	Familiarisation with the Reviews of Marking instructions before the meeting day.	<input type="checkbox"/>

Frequently Asked Questions

Do I need to know how to conduct Reviews of Marking?

No, your task is to just consider the plausible marks that may be awarded to a candidate's response.

Can I write on the scripts?

Yes, you can write on the scripts as you wish.

When I've finished, what should I do with the scripts?

Please could you place them back into the original envelope they were sent to you in and bring them with you to the unit's meeting day with Ofqual in Coventry. The tasks on the meeting day relate to these scripts, so it is very important that you have these with you. Please ensure that the scripts are secure and with you at all times, particularly if travelling via public transport.

Does it matter what order I look at the scripts in?

It does not matter what order you look at the scripts in.

Paper scripts– can I carry out my task on the train/bus home?

Because these are real scripts that have gone through a Review of Marking process they must be treated with confidence, kept securely and marked in a private place.

Why am I undertaking this task?

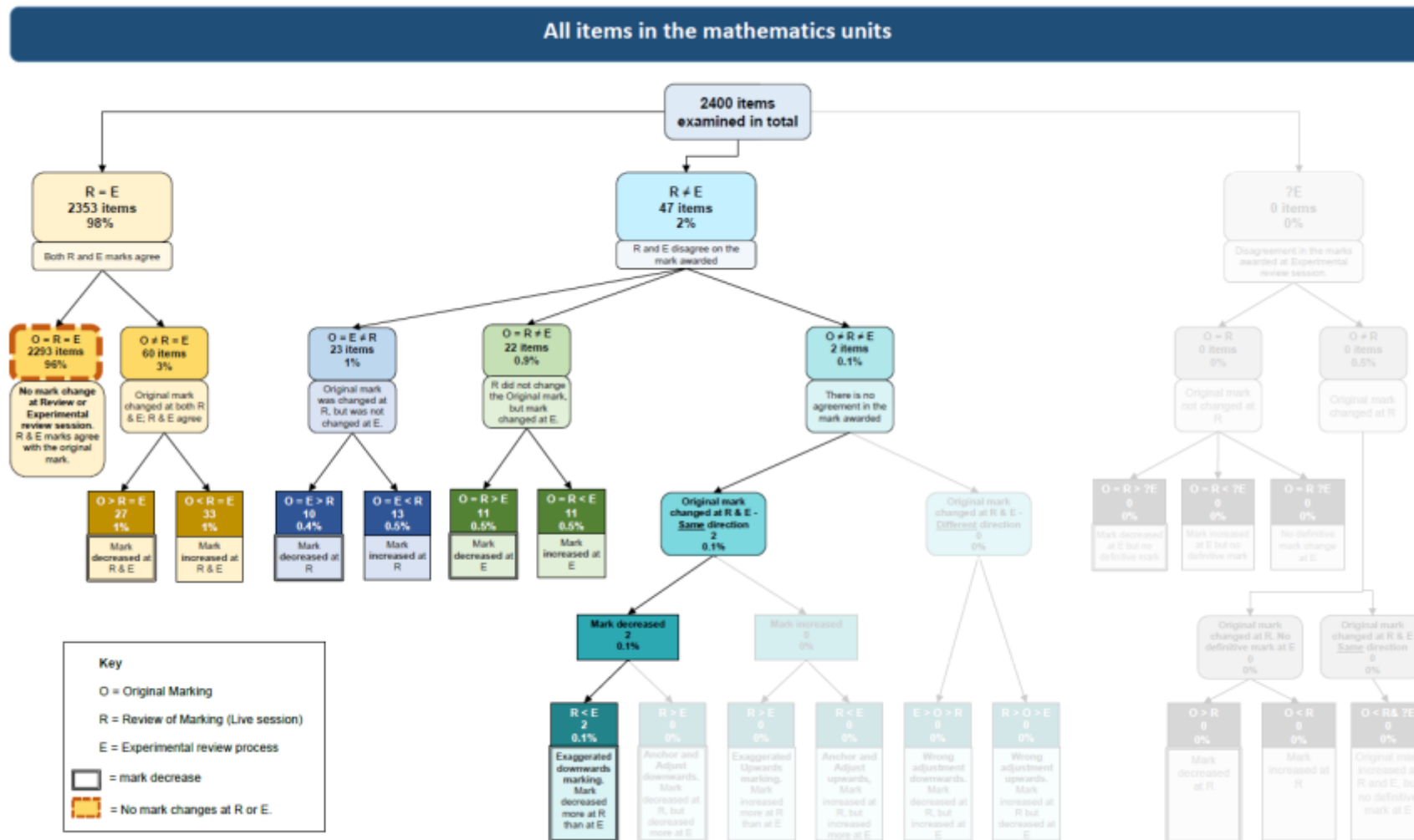
The home-based task directly relates to the tasks you will do during the meeting day. The objective of the meeting day in Coventry is to understand some of the issues and difficulties in conducting Reviews of Marking. Both in the context of your subject, and potentially in other subjects too. We are hoping with some examples of some of the scripts and appropriate expertise in terms of marking and subject expertise, you will be able to contribute to the thinking of this important process.

How do I claim my payment?

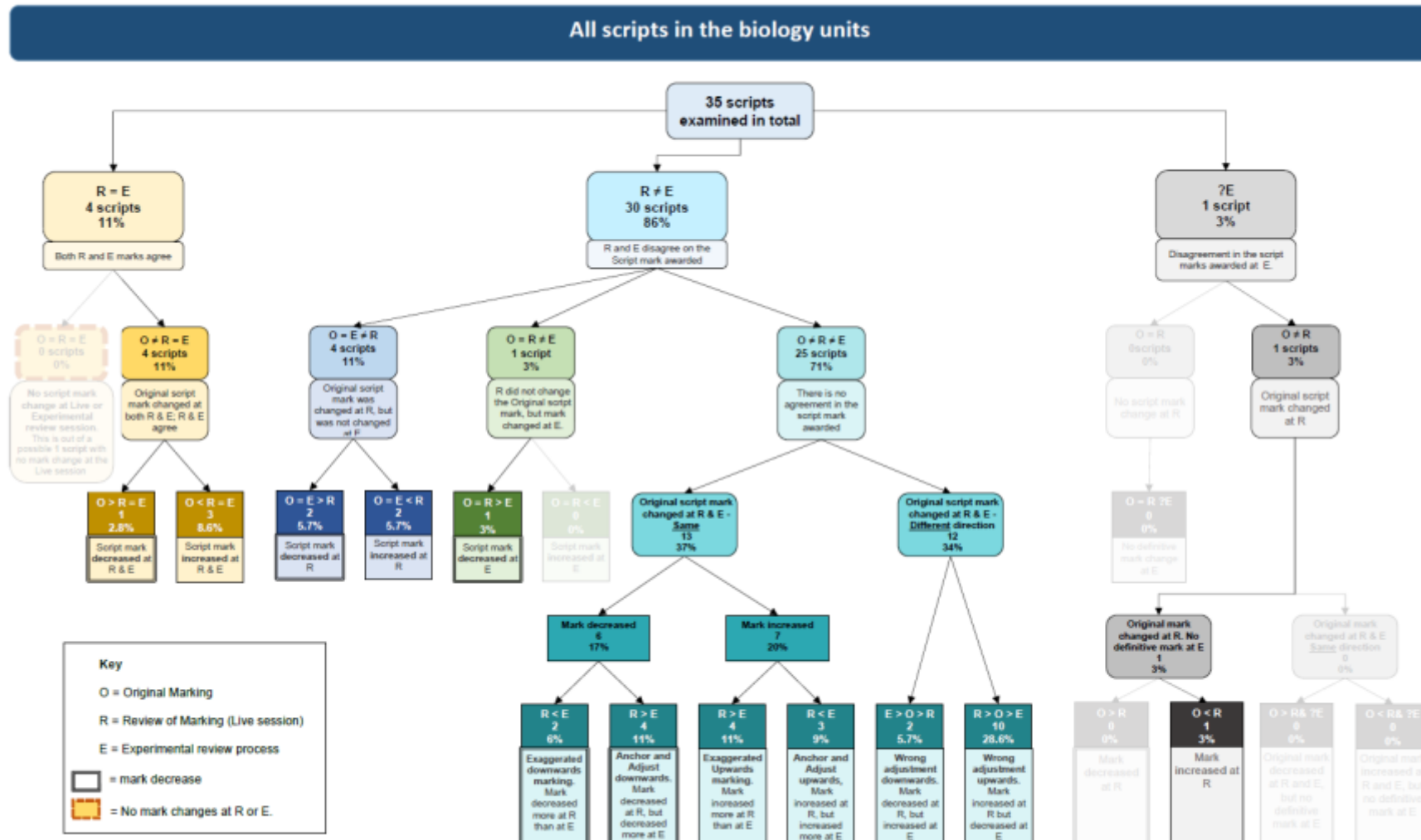
Your payment will be processed after you have completed your work with us, after attending all of your meetings days. You can process this with the finance team following the guidance provided to you in the contract specification.

Who do I contact if I have any questions?

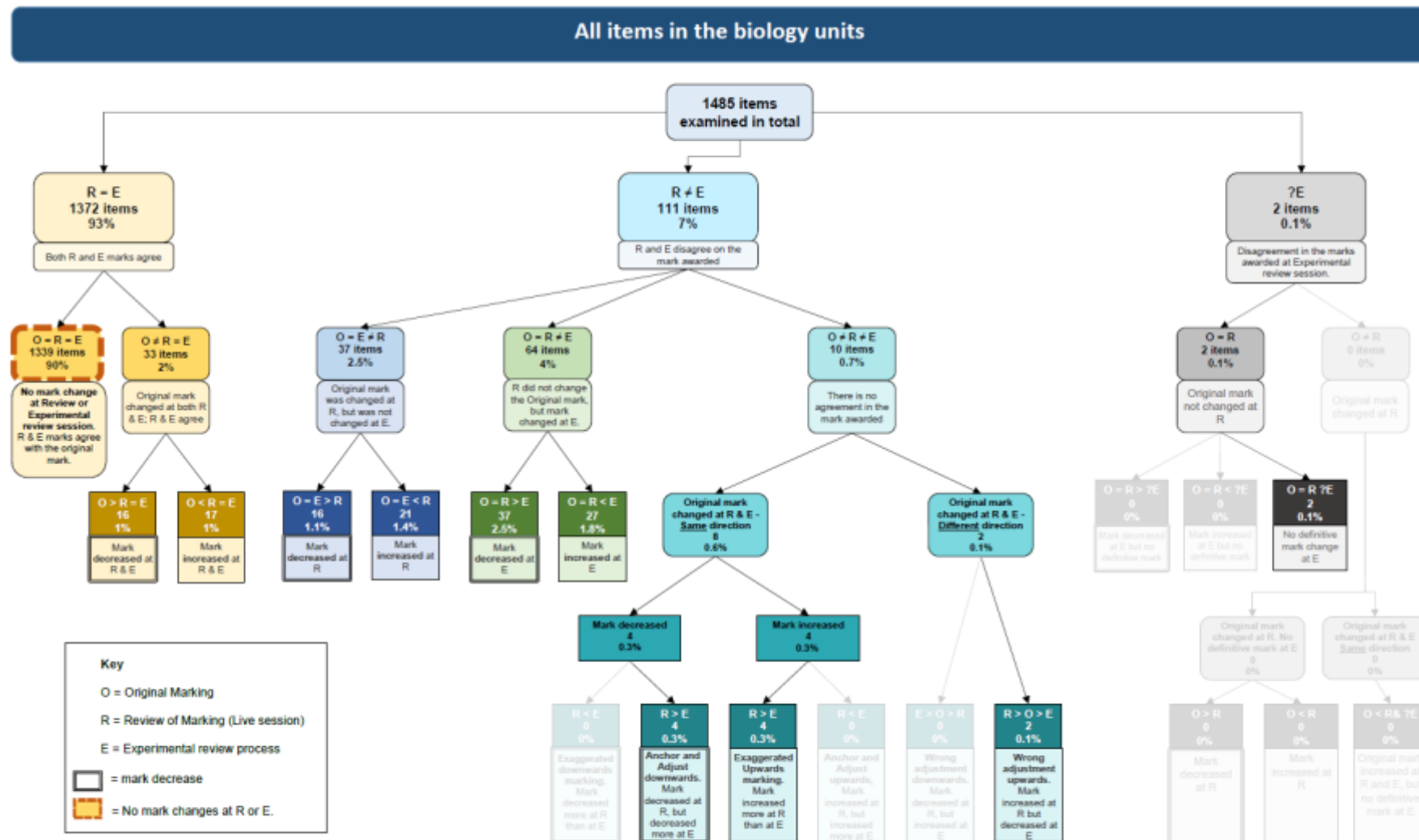
Please contact [name] with any questions you may have:



Note. Numbers indicate the number of items. % indicates the proportion of all items examined. Shaded out areas indicate instances in which mark adjustments were not made at the Review of Marking (Live session) and the Experimental review process. Percentages above 1% are rounded to no decimal places, percentages below 1% are rounded to 1 decimal place.

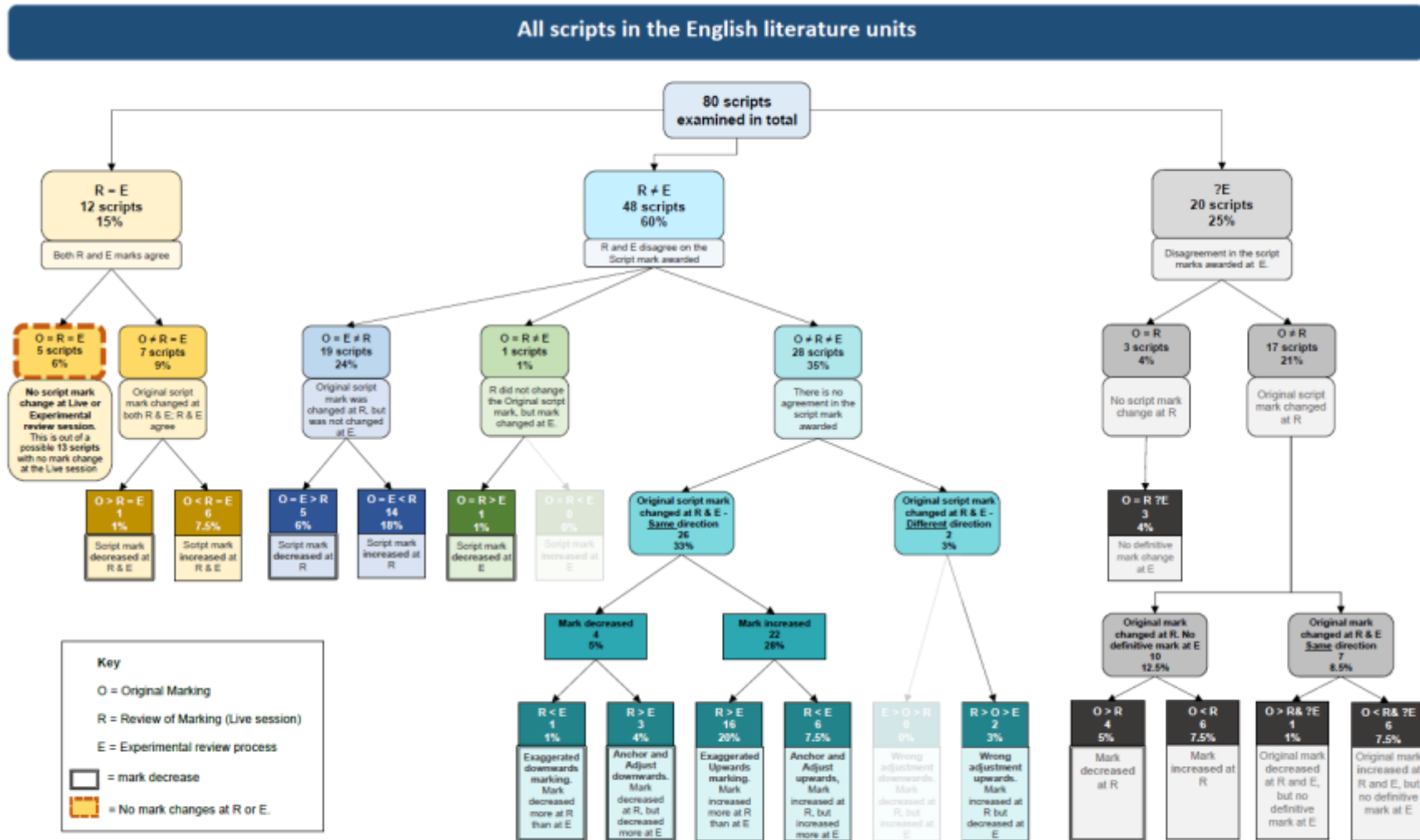


Note. Number indicates the number of items. % indicates the proportion of all scripts examined. Shaded out areas indicate instances in which mark adjustments were not made at the Review of Marking (Live session) and the Experimental review process. Percentages above 1% are rounded to no decimal places, percentages below 1% are rounded to 1 decimal place.

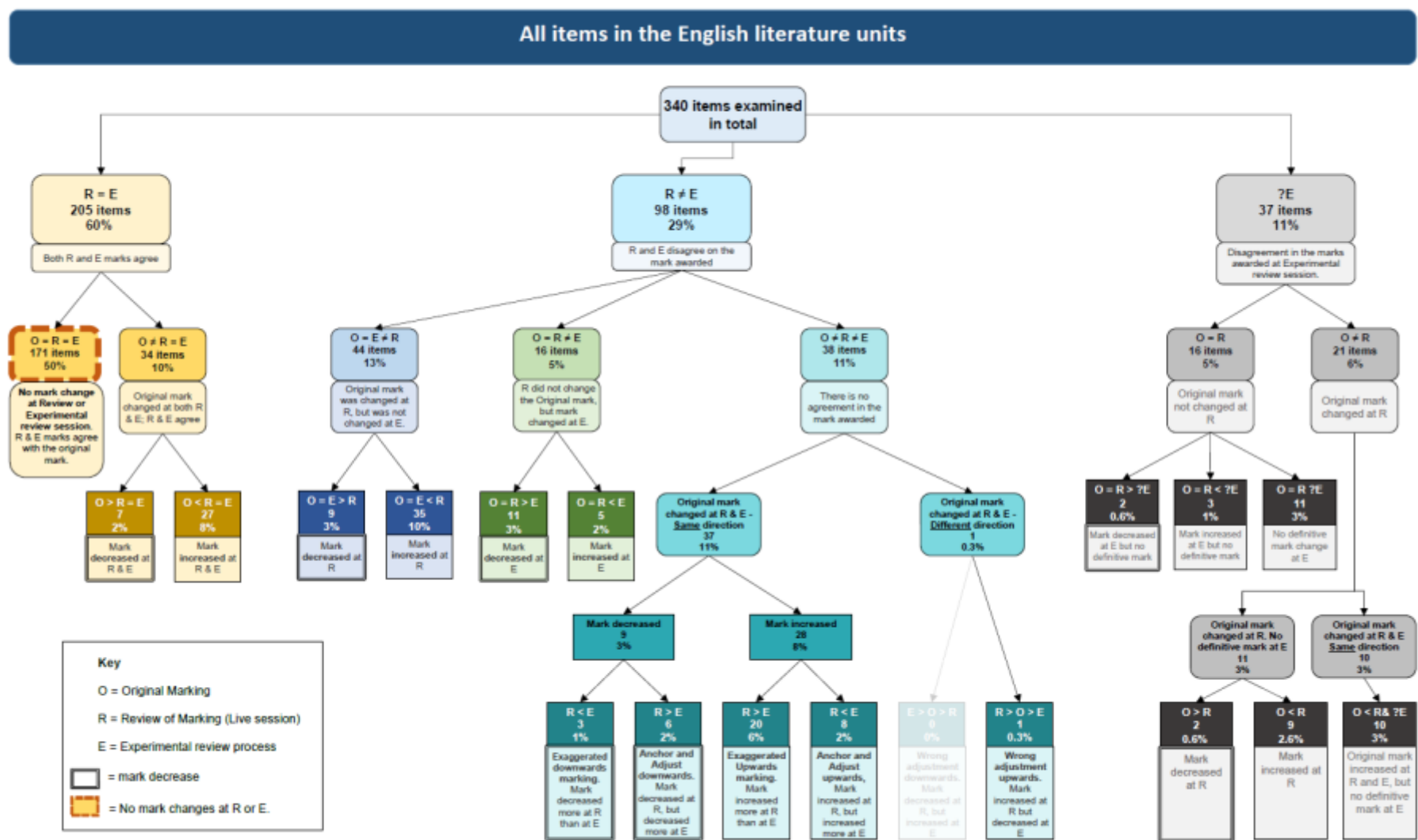


Key
 O = Original Marking
 R = Review of Marking (Live session)
 E = Experimental review process
 □ = mark decrease
 □ = No mark changes at R or E.

Note. Number indicates the number of items. % indicates the proportion of all scripts examined. Shaded out areas indicate instances in which mark adjustments were not made at the Review of Marking (Live session) and the Experimental review process. Percentages above 1% are rounded to no decimal places, percentages below 1% are rounded to 1 decimal place.



Note. Number indicates the number of items. % indicates the proportion of all scripts examined. Shaded out areas indicate instances in which mark adjustments were not made at the Review of Marking (Live session) and the Experimental review process. Percentages above 1% are rounded to no decimal places, percentages below 1% are rounded to 1 decimal place.



Note. Number indicates the number of items. % indicates the proportion of all scripts examined. Shaded out areas indicate instances in which mark adjustments were not made at the Review of Marking (Live session) and the Experimental review process. Percentages above 1% are rounded to no decimal places, percentages below 1% are rounded to 1 decimal place.

Appendix C. The subjects marking and moderation reviewers conducted reviews of marking and moderation on in 2016.

The subjects the marking reviewers conducted reviews of marking on.

Subject	Number of reviewers	Percentage of all subjects
English language/literature	148	11.75%
Mathematics	140	11.11%
Biology	94	7.46%
Geography	90	7.14%
Chemistry	83	6.59%
History	80	6.35%
Physics	72	5.71%
Religious studies	59	4.68%
Business studies	43	3.41%
Psychology	35	2.78%
Economics	32	2.54%
French	29	2.30%
German	26	2.06%
Design and technology	26	2.06%
Sociology	23	1.83%
Science	23	1.83%
Spanish	22	1.75%
Physical education	20	1.59%
ICT	20	1.59%
Other modern languages	19	1.51%
Music	19	1.51%
Computing	16	1.27%
Media/film/TV studies	14	1.11%
Classical subjects	13	1.03%
Health and social care	12	0.95%
Performing/expressive arts	11	0.87%
Political studies	10	0.79%
Law	9	0.71%
Other sciences	8	0.63%
General studies	8	0.63%
Home economics	7	0.56%
Classical subjects	7	0.56%
Statistics	6	0.48%
Drama	6	0.48%
All other subjects	4	0.32%
Travel & tourism	3	0.24%
Citizenship studies	3	0.24%
Welsh	2	0.16%
Other	2	0.16%
Mathematics (further)	2	0.16%
Hospitality	2	0.16%
Critical thinking	2	0.16%
Communication studies	2	0.16%
Art & design	2	0.16%
Welsh literature	1	0.08%
Prep. for life and work	1	0.08%
Leisure & tourism	1	0.08%
Humanities	1	0.08%
Critical thinking	1	0.08%
Additional science (further)	1	0.08%

The subjects the moderation reviewers conducted reviews of moderation on.

Subject	Number of reviewers	Percentage of all subjects
Design and technology	23	12.9%
Art and design	20	11.2%
Science	15	8.4%
ICT	15	8.4%
Music	11	6.2%
History	11	6.2%
English language/literature	10	5.6%
Drama	8	4.5%
All other subjects	7	3.9%
Media/film/TV studies	6	3.4%
Biology	6	3.4%
German	5	2.8%
Health and social care	4	2.2%
Geography	4	2.2%
Business studies	4	2.2%
Physics	3	1.7%
Physical education	3	1.7%
Mathematics	3	1.7%
Home economics	3	1.7%
French	3	1.7%
Engineering	3	1.7%
Computing	3	1.7%
Other modern languages	2	1.1%
Spanish	1	0.6%
Performing/expressive art	1	0.6%
Other sciences	1	0.6%
General studies	1	0.6%
Economics	1	0.6%
Classical subjects	1	0.6%

Appendix D – Reviews of marking survey questions.

In the delivery of the survey there were a number of routing questions indicated by ‘*’.

Question		Response options							
1	Approximately how many Reviews of Marking did you conduct in 2016?								
2	Please list any subjects/unit examinations etc. on which you have conducted Reviews of Marking in the last 3 years and your role								
2.1		Subject:	Art and design Biology Business studies Chemistry Citizenship studies Computing Design and technology Drama Economics Engineering English language/literature	French General studies Geography German Health and social care History Home economics ICT Law Mathematics Media/film/TV studies	Music Performing/expressive art Physical education Physics Political studies Science Social science Sociology Spanish Statistics Other (please state)				
2.2		Level:	GCSE	AS/A level					
2.3	Most senior role on panel:	Assistant Examiner (AE)	Team Leader (TL)	Senior Team Leader (STL)	Assistant Principal Examiner (APE)	Principal Examiner (PE)	Chief Examiner (CE)	Other (please state)	
3	For how many years have you been examining?	<i>Value in years</i>							
4	For how many years have you been conducting Service 2 Reviews of Marking, including when it was formally known as Enquiries About Results?	<i>Value in years</i>							

Evaluation of Reviews of Marking and Moderation 2016 - Study and survey

5	In 2016, were you provided with documentation containing instructions in how to carry out Reviews of Marking? *	Yes	No						
6	In 2016, when did you receive the documentation regarding instructions as how to conduct Reviews of Marking? (Please select all that apply)	June	July	August	September	Other	I'm not sure / I can't remember	If other, please specify	
7	How different were the instructions on the document compared to the previous year (2015)? Instructions were:	...the same as last year	..mainly the same but with a few differences	..quite different from last year	...completely different from last year	I don't know			
8	How important is it to read the instructions document carefully every year?	Essential	Very Important	Reasonably important	Slightly important	Not important			
9	Other than documentation containing instructions, did you receive any training specifically on how to conduct Reviews of Marking in 2016? Examples of any training may be: re-standardisation, online training/briefing, webinar, face to face training/briefing, webinar, face to face training/briefing, etc. *	Yes	No						
10	Please select the type(s) of training you received in how to conduct Reviews of Marking in 2016. (Select all that apply)	Online training	Online briefing	Webinar	Face to face: one to one	Face to face: in a group meeting	Re-standardisation	Face to face briefing	Other (Please specify)
11	When in 2016 did you receive the Review of Marking training? (Please select all that apply)	June	July	August	September	Other	I'm not sure / I can't remember	If other, please specify	
12	In total, how long did it take to complete the training?	Less than 1 hour	1 - 2 hours	2 - 4 hours	4 - 6 hours	6 - 8 hours	Other (please specify)		
13	Thinking about when you conducted Reviews of Marking in 2016, please indicate on the scale the extent to which you agree with the following statements.								

13.1	I felt well prepared to conduct Reviews of Marking:					
13.2	I fully understood how I should conduct Reviews of Marking	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
14	Thinking about when you conducted Reviews of Marking in 2016, please indicate on the scale the extent to which you agree with the following statements.					
14.1	Because of the instructions documentation, I fully understood how I should conduct Reviews of Marking	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
14.2	The training made it clear regarding what I was required to do during a Review of Marking	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
14.3	The training made it clear to me as to the differences between the old and new Review of Marking processes	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
14.4	Because of the training, I felt well prepared to conduct Reviews of Marking	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
15	When conducting Reviews of Marking in 2016, how often did you come across the following circumstances? *					
15.1	The original examiner did not add the question totals or marks correctly	Very often	Often	Sometimes	Occasionally	Not at all
15.2	The original examiner did not see or credit part of an answer	Very often	Often	Sometimes	Occasionally	Not at all
15.3	The original examiner overlooked a credit worthy aspect	Very often	Often	Sometimes	Occasionally	Not at all
15.4	The original examiner misapplied the mark scheme	Very often	Often	Sometimes	Occasionally	Not at all
15.5	The original examiner did not give 'benefit of the doubt' where I would have	Very often	Often	Sometimes	Occasionally	Not at all

15.6	I interpreted differently to the original examiner how the mark scheme should be applied	Very often	Often	Sometimes	Occasionally	Not at all
15.7	I inferred differently to the original examiner a particular word (or part of a diagram) in the response which was difficult to interpret	Very often	Often	Sometimes	Occasionally	Not at all
16	In these circumstances, how likely were you to change the marks?					
16.1	The original examiner did not add the question totals or marks correctly	Definitely change the mark	Very likely	Likely	Unlikely	Definitely not change the mark
16.2	The original examiner did not see or credit part of an answer	Definitely change the mark	Very likely	Likely	Unlikely	Definitely not change the mark
16.3	The original examiner overlooked a credit worthy aspect	Definitely change the mark	Very likely	Likely	Unlikely	Definitely not change the mark
16.4	The original examiner misapplied the mark scheme	Definitely change the mark	Very likely	Likely	Unlikely	Definitely not change the mark
16.5	The original examiner did not give 'benefit of the doubt' where I would have	Definitely change the mark	Very likely	Likely	Unlikely	Definitely not change the mark
16.6	I interpreted differently to the original examiner how the mark scheme should be applied	Definitely change the mark	Very likely	Likely	Unlikely	Definitely not change the mark
16.7	I inferred differently to the original examiner a particular word (or part of a diagram) in the response which was difficult to interpret	Definitely change the mark	Very likely	Likely	Unlikely	Definitely not change the mark
17	Please indicate on the scale the extent to which you agree with the following statements.					
17.1	When conducting Reviews of Marking, I view the process as a review of the original marking: if the original mark could be justified, I do not change it	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
17.2	When conducting Reviews of Marking, I view the process as a review of the	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree

	original marking: if the original mark could be justified, I do not change it					
17.3	In the Reviews of Marking process, I review each response carefully to make sure the original examiner has not missed anything	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
17.4	When conducting Reviews of Marking, I have it in mind that the mark I give should be the same as the original mark	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
17.5	When conducting Reviews of Marking, I view the process as a re-mark in that I will mark everything again and the candidate will always receive my mark rather than the original mark	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
17.6	I think the Reviews of Marking process is mainly confirmatory: I am looking to confirm the original marks	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
17.7	In Reviews of Marking, I try to find a few marks for the candidate	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
17.8	When conducting Reviews of Marking, I believe the mark I give should be different from the original mark	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
17.9	In Reviews of Marking, when the original mark is justified I believe that my mark should still override the original mark	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
18	How does the time you spend per script in Reviews of Marking compare to the time spent on your own prime marking as an original examiner? I spend...:	a lot longer on Reviews of Marking	a bit longer on Reviews of Marking	the same time on both	a bit longer on prime marking	a lot longer on prime marking
19	How do you think the quality of your marking as a Reviews of Marking examiner compares to the marking of the original examiner? My marking is ...:	a lot better	a bit better	the same	a bit worse	a lot worse

20	In general, how easy do you find it to understand how the original examiner has justified their marks?	Very easy	Easy	Neither easy not difficult	Difficult	Very difficult
21	Please indicate the type of mark scheme the question papers on which you conduct Reviews of Marking consist of.*	Only Levels-Based marking	Only Points-Based marking	Both Levels-Based and Points-Based marking		
Points-Based mark schemes						
22	Imagine the following scenarios you may come across for Points-Based mark schemes. Please indicate on the scale the level of agreement your behaviour as a Reviews of Marking examiner has with them.					
22.1	The original examiner gave a response a mark of 3, and I would have ordinarily given it a 4. I think the both of us were right in our interpretation to give the mark. I will keep the mark at a 3.	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
22.2	The original examiner gave a response a mark of 3, and I would have ordinarily given it a 4. I think the original examiner missed a credit worthy point. I will award the mark and give it a 4.	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
22.3	The original examiner gave a response a mark of 3, and I would have ordinarily given it a 2. I think the both of us were right in our interpretation to give the mark. I will leave it as a 3.	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
22.4	The original examiner gave a response a mark of 3, and I would have ordinarily given it a 2. I think the original examiner credited a response that they should not have. I will change the mark and give it a 2.	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Levels-Based mark schemes						

23	Imagine the following scenarios you may come across for Levels-Based mark schemes. Please indicate on the scale the level of agreement your behaviour as a Reviews of Marking examiner has with them.					
23.1	The original examiner gave a response a mark of X. I would have ordinarily given a higher mark but in the same level/band. I will leave the mark as X.	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
23.2	The original examiner gave a response a mark of Y. I would have ordinarily given a lower mark but in the same level/band. I will leave the mark as Y.	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
23.3	The original examiner gave a response a mark of W. I would have ordinarily given a higher mark but in a different level/band. I will leave the mark as W.	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
23.4	The original examiner gave a response a mark of Z. I would have ordinarily given a lower mark but in a different level/band. I will leave the mark as Z.	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree

Appendix E – Reviews of moderation questions and response options.

In the delivery of the survey there were a number of routing questions, the details of which are not included her.

Question		Response options																																
1	Approximately how many Reviews of Moderation did you conduct in 2016?																																	
2	Please list any subjects/unit examinations etc. on which you have conducted Reviews of Moderation in the last three years and your role	Subject:																																
2.1		Art and Design	Biology	Business Studies	Chemistry	Citizenship Studies	Computing	Design and Technology	Drama	Economics	Engineering	English Language / Literature	French	General Studies	Geography	German	Health and Social Care	History	Home Economics	ICT	Law	Mathematics	Media / Film / TV studies	Music	Performing / Expressive Art	Physical Education	Physics	Political Studies	Science	Social Science	Sociology	Spanish	Statistics	Other (please state)
2.2		Level:	GCSE	AS/A level																														
2.3		Most senior role on panel:	Moderator	Team Leader Moderator	Assistant Principal Moderator	Principal Moderator	Chief Moderator	Other (please state)																										
3	For how many years have you been a moderator?	Value in years																																
4	For how many years have you been conducting Reviews of Moderation/Service 3 Enquiries About Results?	Value in years																																
5	In 2016, were you provided with documentation containing instructions	Yes	No																															

Evaluation of Reviews of Marking and Moderation 2016 - Study and survey

	in how to carry out Reviews of Moderation? *								
6	In 2016, when did you receive the documentation regarding instructions as how to conduct Reviews of Moderation (Please select all that apply)	June	July	August	September	Other	I'm not sure / I can't remember	If other, please specify	
7	How different were the instructions on the document compared to the previous year (2015)? Instructions were:	...the same as last year	..mainly the same but with a few differences	..quite different from last year	...completely different from last year	I don't know			
8	How important is it to read the instructions document carefully every year?	Essential	Very Important	Reasonably important	Slightly important	Not important			
9	Other than documentation containing instructions, did you receive any training specifically on how to conduct Reviews of Moderation in 2016? Examples of any training may be: re-standardisation, online training/briefing, webinar, face to face training/briefing, webinar, face to face training/briefing, etc.*	Yes	No						
10	Please select the type(s) of training you received in how to conduct Reviews of Moderation in 2016. (Select all that apply)	Online training	Online briefing	Webinar	Face to face: one to one	Face to face: in a group meeting	Re-standardisation	Face to face briefing	Other (Please specify)
11	When in 2016 did you receive the Review of Moderation training? (Please select all that apply)	June	July	August	September	Other	I'm not sure / I can't remember	If other, please specify	
12	In total, how long did it take you to complete the training?	Less than 1 hour	1 - 2 hours	2 - 4 hours	4 - 6 hours	6 - 8 hours	Other (please specify)		
13	Thinking about when you conducted Reviews of Moderation in 2016, please indicate on the scale the extent to which you agree with the following statements.								

13.1	I felt well prepared to conduct Reviews of Moderation	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
13.2	I fully understood how I should conduct Reviews of Moderation	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
14	Thinking about when you conducted Reviews of Moderation in 2016, please indicate on the scale the extent to which you agree with the following statements.					
14.1	Because of the instructions documentation, I fully understood as a reviewing moderator how I should conduct Reviews of Moderation	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
14.2	The training made it clear regarding what I was required to do during a Review of Moderation	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
14.3	The training made it clear to me as to the differences between the old and new Review of Moderation processes	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
14.4	I felt well prepared to conduct Reviews of Moderation as a result of the training	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
15	What information are you provided with when conducting a Review of Moderation? Please select all that apply. *					
15.1	Centre number					
15.2	Centre name					
15.3	Names of candidates					
15.4	The candidate's work					
15.5	The mark scheme					
15.6	The original moderator's name					

15.7	The teacher's marks					
15.8	The original moderator's marks					
15.9	The original moderator's rationale for any individual marks					
15.1	The original moderator's report on the centre					
15.11	The adjustment that has been applied to the centre's marks.					
16	When conducting a Review of Moderation, how important is each of these sources of information for influencing your judgements?					
16.1	Centre number	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
16.2	Centre name	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
16.3	Names of candidates	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
16.4	The candidate's work	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
16.5	The mark scheme	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
16.6	The original moderator's name	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
16.7	The teacher's marks	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
16.8	The original moderator's marks	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
16.9	The original moderator's rationale for any individual marks	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
16.1	The original moderator's report on the centre	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant

16.11	The adjustment that has been applied to the centre's marks.	Very important	Important	Neither important nor unimportant	Unimportant	Very unimportant
17	Imagine the following scenarios in which you are looking at an individual candidate's work in a Review of Moderation. Please indicate which mark you would record.					
17.1	The teacher mark is X, the original moderator mark is lower than X (but in tolerance). You agree more with the teacher's mark.		I would record the teacher's mark		I would record the original moderator mark	
17.2	The teacher mark is X, the original moderator mark is lower than X (but out of tolerance). You agree more with the teacher's mark.		I would record the teacher's mark		I would record the original moderator mark	
17.3	The teacher mark is X, the original moderator mark is higher than X (but in tolerance). You agree more with the teacher's mark.		I would record the teacher's mark		I would record the original moderator mark	
17.4	The teacher mark is X, the original moderator mark is higher than X (but out of tolerance). You agree more with the teacher's mark.		I would record the teacher's mark		I would record the original moderator mark	
18	Think about when you conduct a Review of Moderation on a set of centre work. Please indicate the extent to which you agree with the following statements.					
18.1	I review each candidate's piece of work entirely independently of the others from the centre	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree

18.2	I am expecting to be able to confirm the original moderator's set of marks for the centre	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
18.3	I am looking for a pattern in the relationship between the centre's and moderator's mark	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
18.4	I am trying to work out whether I can agree with the original moderator's marks	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
18.5	After I have looked at 3 or 4 pieces of candidate work, I get a sense of how accurate the centre's marking is	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
18.6	After I have looked at 3 or 4 pieces of candidate work, I get a sense of how accurate the original moderator's judgments are	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
19	Have you ever changed an overall moderator's decision?	Yes	No	I don't know		
20	Please describe the circumstances in which you have changed an original moderator's decision.	<i>Free writing box</i>				
21	Please indicate the type of mark scheme which you use for moderation / Reviews of Moderation.	Only Levels-Based marking	Only Points-Based marking	Both Levels-Based and Points-Based marking		
22	Describe how you use tolerance when conducting a Review of Moderation	<i>Free writing box</i>				
23	Sometimes during Reviews of Moderation you may want to give a mark which is out of tolerance. Please indicate on the scale your level of agreement with the following statements.					
23.1	If I feel it is the right mark based on the candidate's work, I will give a new mark which is out of tolerance	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree

23.2	If there are already one or more marks out of tolerance, I feel more confident about giving a mark out of tolerance	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
23.3	It depends on the overall centre marking. If the centre marking is generally fine, I will not record a mark out of tolerance	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
23.4	If there are already one or more marks out of tolerance, I may feel cautious about giving a mark out of tolerance	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree

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



© Crown copyright 2017

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit <http://nationalarchives.gov.uk/doc/open-government-licence/version/3> or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or email: publications@ofqual.gov.uk.

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.

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

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