Research and Analysis

Marking consistency studies

Summer 2016 and 2017 units

Authors

This report was prepared by Stephen Holmes, Caroline Morin, Ben Cuff and Beth Black from Ofqual's Strategy, Risk and Research Directorate.

Acknowledgements

We would like to thank the awarding organisations for providing scripts and materials, and also thank the examiners who took part in the studies, particularly the Principal Examiners.

Contents

Executive summary

Ofqual published a report in February 2014 reviewing the quality of marking in GCSEs and A levels which included a commitment to better monitor and quantify the accuracy of this marking. At the Ofqual Summer Series Symposium in June 2016, we announced the start of a programme of 2 parallel strands quantifying marking consistency. The marking consistency metrics strand used live monitoring data from the exam boards to measure marking consistency. However, publishing metrics for individual components could compromise the live monitoring procedures. Therefore the second strand involved carrying out our own marking consistency studies, which are reported here. These studies allow us to analyse specific units, as well as helping us understand the overall validity of the marking consistency metrics.

This report describes 2 rounds of marking consistency studies on a range of summer 2016 and summer 2017 units. Examiners who had marked on these units in the summer series were recruited and each marked the same set of scripts within each unit. This multiple marking of items allowed us to measure the variability in the marks awarded, understand where there are similarities and differences in marking consistency between and within subjects, and publish accordingly.

For 23 units (14 from 2016 and 9 from 2017), in 7 subjects, a team of up to 8 examiners per paper were recruited according to their role in the live summer series. This team usually comprised a principal examiner (in charge of marking on the paper), 2 team leaders and 5 assistant examiners. They all marked 100 full scripts on a bespoke online marking system.

The analysis of marking consistency was carried out at both whole script level (the mark assigned to each candidate script by each marker) and at individual item level. The mean difference to the principal examiner's mark (the 'definitive' mark) and the standard deviation (spread) of the differences was calculated. The mean and standard deviation of each marker to the median mark (the consensus mark) was also calculated.

There were substantial differences in the spread of whole script mark differences across the units, indicating that marking showed varying levels of consistency across units and subjects. The spread of the whole script mark differences were a little larger here than was observed in the marking consistency metrics, for several reasons, likely including the types of candidate responses included and the lower-stakes nature of this marking exercise compared to live summer marking. However, the correlation across units between the measures of script mark differences here and in the marking consistency metrics was very high, indicating that the 2 different ways of collecting data are both valid methods for understanding marking consistency.

1 Introduction

In February 2014, Ofqual published a report on the quality of marking in GCSEs and A levels[1]. In this report, we pledged to lead a programme of improvements and identified 6 steps that should be taken to improve the quality of marking. One of these steps was to better monitor and quantify the quality of marking of general qualifications. Following this report, at our Summer Series Symposium in June 2016[2] we committed to carrying out a rolling programme covering both analysis of the exam boards' own monitoring data (marking consistency metrics), and our own marking consistency studies.

The marking consistency metrics[3] use data from live marking monitoring provided by the exam boards to generate measures of marking consistency. Seed items used in monitoring are pre-marked by the principal examiner and sometimes a small group of senior examiners, to give a 'definitive' mark. They are then introduced into each marker's allocation at intervals and used to monitor whether that examiner is marking correctly to the standard. We do not want to compromise these live monitoring procedures by publishing the metrics for individual components, so by carrying out our own marking studies we are able to look at specific units and also help us understand the overall validity of the marking consistency metrics.

This report describes the 2 rounds of marking consistency studies that we have carried out to date looking at units (components) from the 2016 and 2017 summer series. In January 2017 and January 2018, we requested material from the exam boards in order to carry out these studies. The aim was to quantify marking consistency across the 4 exam boards in 14 units from the summer 2016 series and 9 units from the summer 2017 series. For all units, a number of examiners and senior examiners each marked the same 100 clean (ie anonymised, and free of any annotations) scripts, and the marks they awarded were compared to a) the marks given to these scripts by the Principal Examiner (the 'definitive mark') and b) the median of the marks awarded by the other examiners (the 'consensus mark').

We can compare the measures of consistency between markers calculated in this study to those generated by the marking consistency metrics. If the 2 sets of measures of marking consistency are in good agreement, this indicates that the 2 different ways of collecting data are both valid methods for understanding marking consistency. In other words, we can have confidence in both methods as ways of understanding the quality of marking.

2 Methods

2.1 Units selected and recruitment

We first selected the subject/level combinations we wished to include in each round of the study. Subject choice was based on obtaining a good sample of different

[1] https://www.gov.uk/government/publications/quality-of-marking-in-gcses-and-a-levels
[2] https://www.gov.uk/government/news/summer-series-symposium-29-june-2016
[3] https://www.gov.uk/government/publications/marking-consistency-metrics and https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf

types of questions, from low to high tariff, and with points-based and levels of response mark schemes. For the 2016 study, we chose units in reformed and non-reformed subjects that had a high number of reviews of marking (if there is a concern that an error or misapplication of the mark scheme has occurred, candidates can request the exam board to review the marking). This was because the 2016 study was carried out in conjunction with another study looking at reviews of marking. For the 2017 study, individual units were picked that had similar content and examination length across exam boards. For both years, the selection of exam boards for each subject was based on the numbers of candidates, and indirectly the number of examiners (so as to have a sufficient pool of examiners from which to recruit). Table 1 presents the units used in both studies, as well as the number of examiners recruited for each unit.

We provided the exam boards with recruitment emails to send to suitable examiners. Some exam boards chose to select the examiners themselves and provided us with a list of contact details, while others sent out batch emails to the full examiner team and requested that they reply directly to Ofqual; we then recruited to each role on a first-come first-served basis. For each unit, where possible, a principal examiner (PE), 2 team leaders (TL) and 5 assistant examiners (AE) were recruited. The intention was to provide a representative cross-section of the examiner population. Sometimes we were unable to achieve this target, as shown in Table 1. In 2 cases the PE was unavailable when the study was due to take place and so the Assistant Principal Examiner (APE) was recruited and carried out the PE role.

2.2 Materials

We requested the scripts/items that were used during the original, live standardisation so that the same standardisation materials could be used in the study. This included any (annotated) practice scripts with definitive marks for examiners to review before starting on the standardisation (sometimes called qualification) scripts, as well as the definitive marks the PE (and their senior colleagues) had given to the standardisation items. Depending on the process used by the exam board for standardisation, only some units included annotated practice scripts.

For each unit, we used the full set of script marks to select a sample of 130 scripts that had a mean and range of marks matching the full set. Scripts receiving less than 5 marks in live marking were not included, as they were likely to contain very little credit-worthy material (meaning that they would provide little insight into marking variability). The scripts for these samples were requested from the exam boards. For the 2016 study we did not request anonymised scripts, and so we anonymised them ourselves, removing any identifying features. For the 2017 study we asked the exam boards to carry out this anonymisation themselves.

Table 1: *Summary of examiners who were recruited for the 2016 and 2017 studies and numbers whose data was included and excluded in the final analysis.*

Year / subject	Exam board	Examiners recruited			Examiners included in analysis			Examiners excluded from analysis
		PE	TL	AE	PE	Tl	AE	All
2016								
Biology GCSE	AQA	1	2	5	1	2	5	-
	OCR	1	2	5	1	2	5	-
	Pearson	1	2	4	1	2	3	1
English language GCSE	AQA	1	2	5	1	2	4	1
	OCR	1	2	4	1	1	3	2
	Pearson	1	2	6	1	2	4	2
	WJEC	1	2	5	1	2	3	2
English literature AS	AQA	1	2	4	1	2	4	-
	OCR	1	2	5	1	2	3	2
	Pearson	1	2	5	1	1	2	4
	WJEC	1	1	3	1	1	1	2
History AS	AQA	1	2	5	1	2	4	1
	OCR	1	2	3	1	2	3	-
	Pearson	1	2	5	1	2	5	-
2017								
Economics A level	AQA	1	2	3	1	2	2	1
	Pearson	1	2	4				-
English literature GCSE	AQA	1	2	5	1	2	4	1
	OCR	1	2	4	1	2	4	-
	Pearson	1	2	5	1	1	4	2
	WJEC	1	2	5	1	2	2	3
Mathematics GCSE	AQA	1	2	5	1	2	5	-
	OCR	1	2	5	1	2	5	-
	Pearson	1	2	5	1	2	5	-

From the larger sample of 130 scripts, 100 were selected to be used in the study. We initially excluded scripts that either used a scribe or included additional answer booklets or pages, as these were more likely to lead to responses being mis-marked on our marking system, and/or would often be removed from the allocations given to examiners in live marking. Where the remaining sample was above 100, it was reduced to 100 through random selection. If the remaining sample fell below 100, some of these atypical scripts were included, particularly where the candidate had clearly indicated the continuation or the layout of the responses on the script was unlikely to lead examiners to miss parts of the response.

2.3 Marking software

We commissioned a bespoke marking system from an external supplier. The main advantage of the bespoke system was that it should be a system new to all examiners, and not potentially confer advantage to one board over others by way of examiner familiarity. In other words, it should introduce no bias in our data for particular units.

Each exam board uses its own online marking system, and these are configured to allow marking in slightly different ways in respect of whether candidate work is marked by item or by script. To allow all our recruited examiners to mark in their usual way, the system was designed to be flexible, to allow marking to take place both by item (for example marking Q1 for all scripts, followed by Q2 for all scripts) and by whole script (marking all items within a script, before moving on to the next script) or a combination of the two. In the 2016 study the system allowed navigation through the items along both dimensions within each marking session, but for the 2017 study the system was designed to force a decision at the start of each marking session, to either mark through individual items, or down individual scripts. In both cases we instructed the examiners to mark as they would mark on their exam board's own system. We did not force this; – the final choice of method to use was the examiner's own.

The system was generally found to be intuitive to use (we surveyed all the examiners on their use of the system at the end of the exercise), although we provided detailed instructions, both written and in the form of video tutorials covering all the different methods of marking through items. Some units had defined response areas on the question papers, and in these instances the system moved to the relevant part of the script for each question. Unlike some systems used by the exam boards, it did not just display the 'clip' – the pre-set area defined for each answer – but allowed full scrolling up and down on the script. This meant that if a candidate wrote outside the normal response area (or on a continuation booklet at the end) the answer was not cut off. For other units, which used unstructured answer booklets, the examiners had to locate each item themselves on the full scripts.

One difference (for 2016 only) between the bespoke system used in this study and the systems that the exam boards use was that there was no access to annotation tools. The bespoke system was updated in 2017 and annotations were available to examiners who marked in the 2017 study. These were a combination of pre-defined annotations (taken from the mark schemes and/or agreed with the PEs) and open comments.

2.4 Procedure

In both years the marking took place during February and March, around 8 months after the live summer marking window. To remind the examiners of the standard they should be marking to, we revisited the standardisation materials that had been used in the summer. Prior to main marking, the examiners first reviewed any practice scripts we were sent and then they marked the standardisation scripts online. The marks for these were then sent to the PE. For the 2017 study, copies of the examiners' annotations on the scripts were also sent to the PE. The PE then compared the marks awarded by the examiners to their definitive mark. Where required, we asked the PEs to have discussions with the examiners (via email and/or over the phone) to ensure that their marking was similar to the standard required during live marking. Once this was done, and the PE was satisfied with their marking, the examiners were allowed to start marking the 100 scripts. Standardisation and marking took place over a 2-week marking window. In the 2017 study, examiners were asked to annotate responses in the way they had in the summer marking, to encourage them to mark more like they had done the previous summer.

Each examiner, including the PE, marked the same 100 scripts in a random order to avoid sequence effects. When marking by item, the same random script order was used within each item (but this order was different for each examiner). There were no restrictions on the number of scripts or responses that an examiner could mark in a session. Examiners were asked to email the researchers when they had completed marking all scripts. A short questionnaire was then sent to the examiner to complete.

Unlike the marking systems used by the exam boards, the bespoke system did not allow for the monitoring of examiners' performance during marking using seed items (items pre-marked by the senior examiners which are inserted into each examiners' marking allocation) or backreading (re-marking of a sample of an examiner's scripts/items by a more senior marker). Instead, at the end of the marking exercise all of the PEs were given an opportunity to grade all the examiners on their unit by looking at a sample of their marks and applying a rating scale similar to that used at the end of live marking within the exam boards. Ratings went from one (good, only very minor issues) through 2 (generally good with some occasional issues) and 3 (just about adequate) to 4 (problematic).Those who were rated 4 were not included in the analyses. Table 1 details how many examiners were retained or excluded for the final analysis.

It is worth noting that the exclusions made here may be different to those that would have occurred in the monitoring of live marking. Rather than basing exclusions on rules around the closeness of seed marks to the definitive mark or backreading, exclusions are here based upon a holistic view of a random selection of each marker's full scripts, compared to the PE's mark on the same selection. This is also different to the live monitoring of most units that are item-level marked, where monitoring and exclusions are made at individual item level.

3 Results

We consider marker differences from both the PE's 'definitive' mark[4] (as they marked the same set of scripts as the other examiners) and the median (consensus) mark in this analysis. Initially we consider mark differences at whole script level, then we analyse mark differences for individual items grouped by tariff.

It should be noted that both analyses (PE mark and median mark) involve comparing examiners' marks for a script or response to a single mark rather than a range of marks. It is possible, however, especially in subjects like English Literature, that there are is a small number of legitimate marks for some scripts/responses. Thus the marker agreement metrics are relatively stringent in that they do not take this into account. However, they do provide a constant metric for comparison between components within subject groupings.

3.1 Script level differences

This section describes marking accuracy at whole script level. Two benchmarks are used; the difference of each script mark for each examiner from the PE mark and the group median mark (including the PE mark). The former shows the spread of examiner script marks, and also how their mean compares to the PE; were they all matching the PE standard or were they consistently harsher or more lenient? The comparison to the group median mark shows how broadly spread the marks from the whole group including the PE were.

All of our marking was carried out on whole scripts, even where a marker chose to work through their allocation one question at a time, meaning that all scripts were marked in their entirety by each examiner. In the live summer marking, some units are marked as whole scripts but other units are marked at item level, with items from a single candidate script distributed across multiple markers. This means that whole script marks are combined from the marks of several markers. Our estimates of whole script mark difference may be different to that achieved with these distributed items in live marking. Appendix A describes simulations to evaluate the relative size of the mark difference in whole-script marks or distributed-item script marks, which showed that the whole script mark differences are on average only around 10% larger than distributed-item script differences.

3.1.1 Difference from PE mark

The mean and standard deviation of the examiner difference from the whole-script PE mark, scaled as a percentage of the marks available on the paper[5], is shown for all the 2016 units (Figure 1) and all the 2017 units (Figure 2). The PE sets the definitive item mark in this data and is therefore not included in the set of examiners.

[4] The term 'definitive' grade or 'definitive' mark is based on the terminology ordinarily used in exam boards for the mark given by the senior examiners at item level for each seeding response. Thus, although it is possible that there is more than one legitimate mark for some responses, the system does not capture these.

[5] The script totals used to calculate the scaled differences (% of maximum mark) include all items on each unit even though in some cases our markers did not mark every item. Objectively marked items (eg MCQ or diagram completing) are usually auto-marked in live marking (with 0 mark disagreement) and were therefore not marked here.

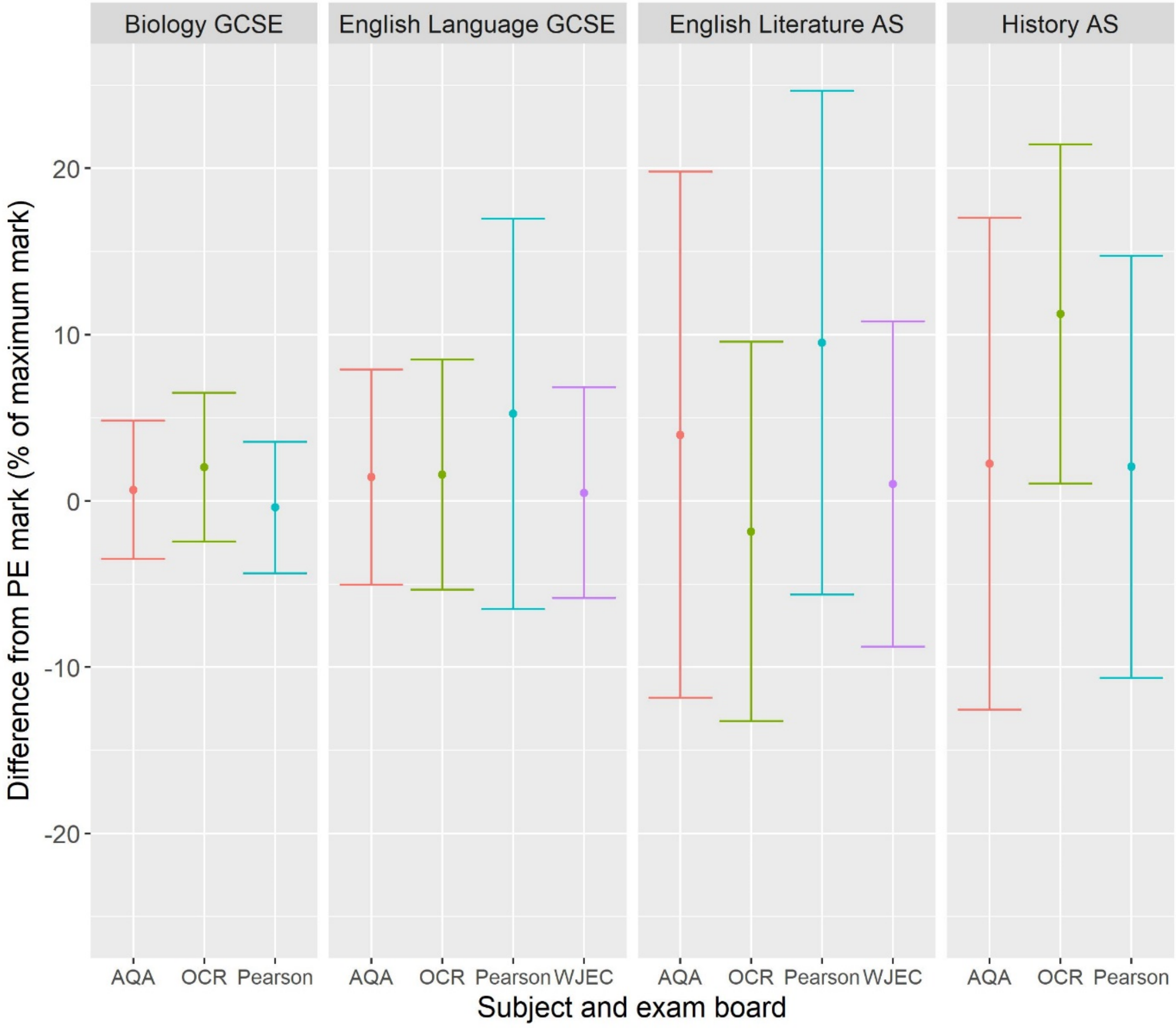


Figure 1: *Mean and standard deviation of whole script mark differences from the PE mark, averaged across all markers and all scripts, for summer 2016 units.*

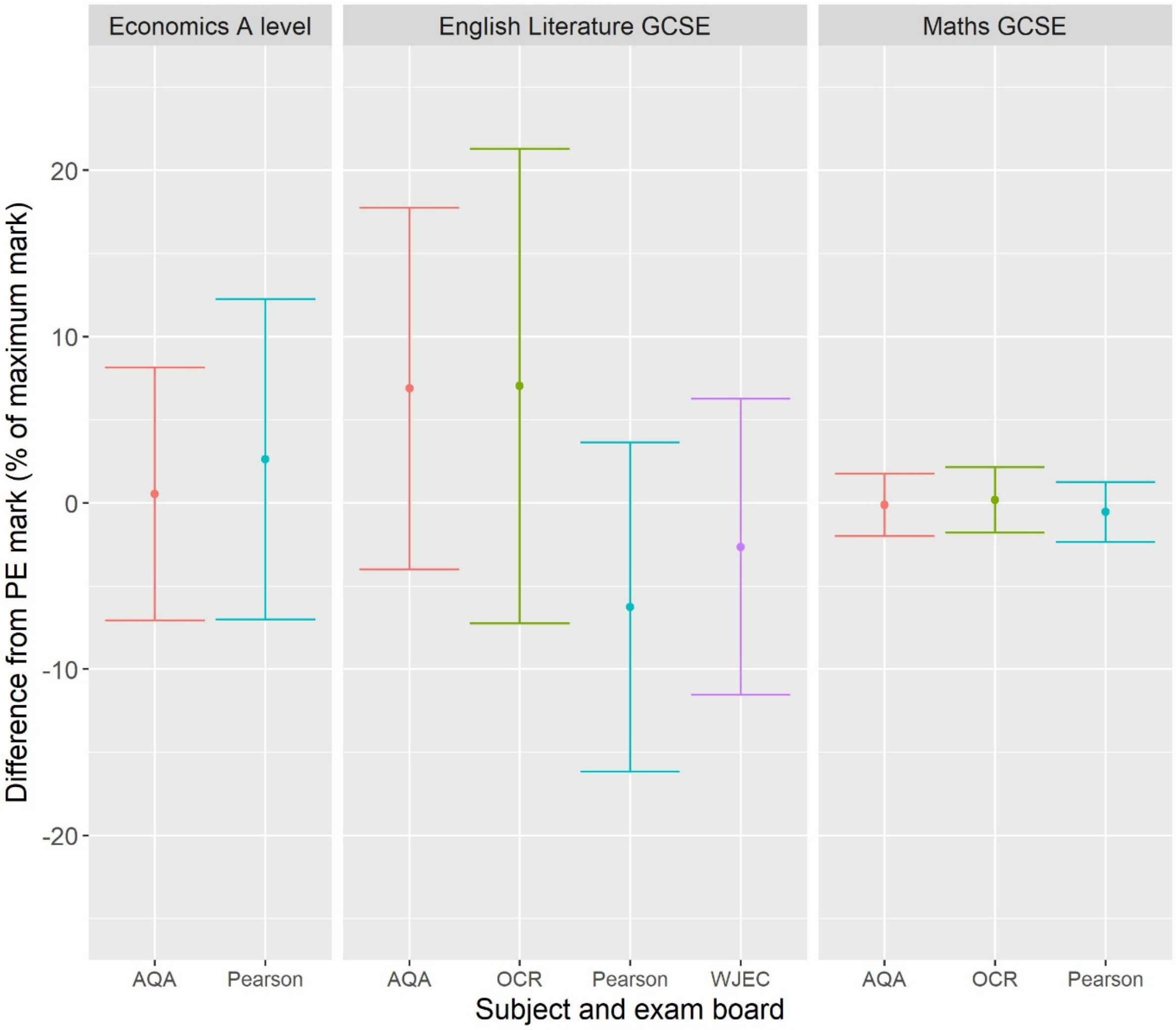


Figure 2: *Mean and standard deviation of whole script mark differences from the PE mark, averaged across all markers and all scripts, for summer 2017 units.*

There are clear differences in the spread of script-level differences both across and within subjects. There are also a number of units where the mean difference was non-zero. This indicates that the examiners were, as a group, marking more leniently (when the mean difference was positive) or more severely (mean difference negative) than the PE. This would suggest that for some reason, perhaps their understanding of the standardisation materials or some other factor in the marking task, the group as a whole were slightly misaligned with the PE's standard. However, it is possible that the PEs themselves marked slightly differently to the standard set in the standardisation materials. Given the length of time that had passed since the PEs selected and marked the standardisation materials, we cannot rule out that their own standard had drifted a little and that we did not allow them sufficient warm-up into this exercise.

To validate the marking consistency measures obtained here, we correlated the standard deviations of the script-level differences (as a proportion of maximum marks) for all of the units in this study where we had corresponding data from the marking consistency metrics[6] work. We obtained a significant Pearson correlation coefficient (r(13) = 0.94, p < 0.001) indicating very strong agreement between the measures of marker consistency from the 2 studies (see Figure 3). The mean standard deviations of the mark differences for each subject across the 2 studies are shown in

[6] See https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf

Table 2. The standard deviations we have obtained in this study are larger, by just over a third on average (36%: marking consistency studies 7.1% vs marking consistency metrics 5.2%).

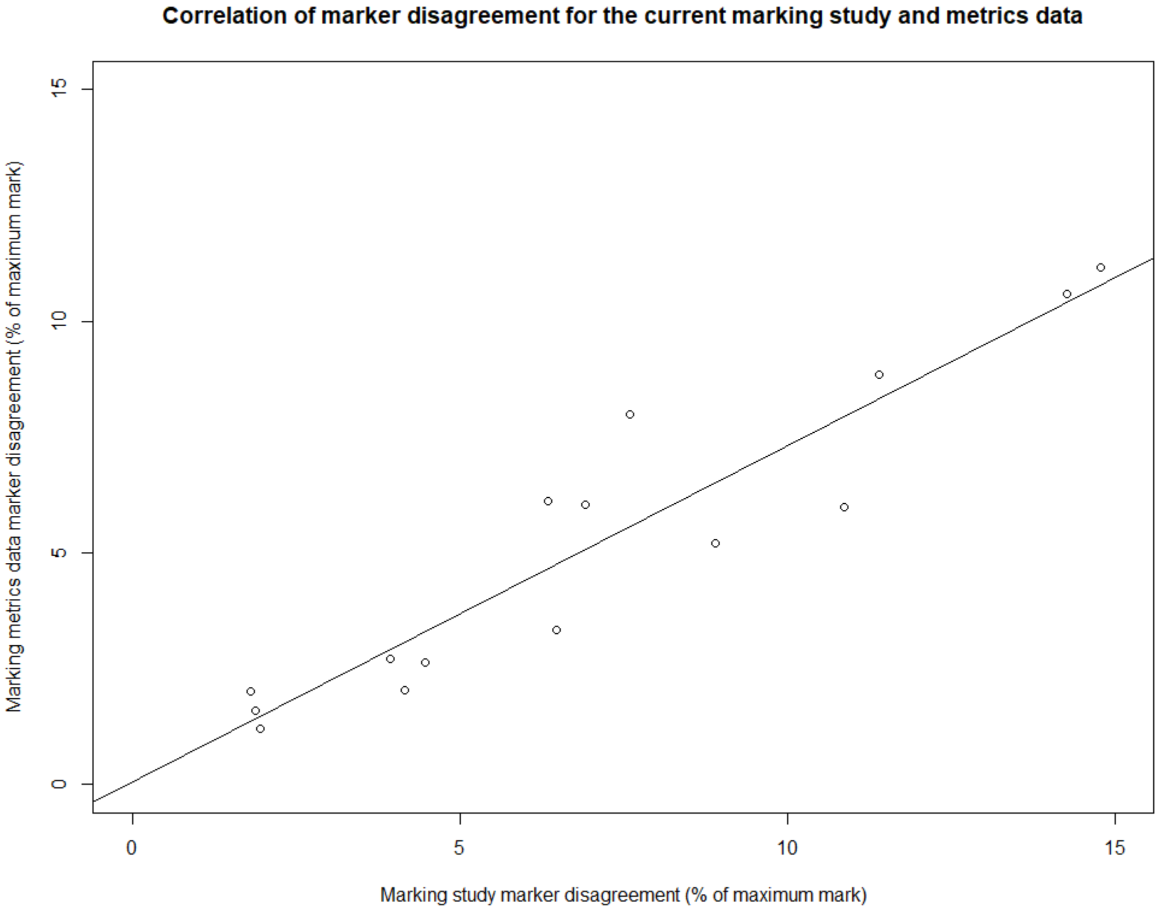


Figure 3: *Scatterplot showing the standard deviation of the whole script marker disagreement obtained for 15 units, from both 2016 and 2017, in the current marking study and the corresponding data from the marking metrics. The regression line is shown.*

Table 2: Mean standard deviation of whole script mark differences as a percentage of maximum marks, combined across units within each subject, for the units that are common to both the marking consistency study and the marking consistency metrics.

Subject	Marking consistency studies (%)	Marking consistency metrics (%)
Biology GCSE	4.2	2.5
English Language GCSE	6.6	5.2
English Literature AS	11.4	8.9
History AS	14.8	11.2
Economics A level	7.6	8.0
English Literature GCSE	11.3	7.3
Mathematics GCSE	1.9	1.6

Appendix B contains distributions of script-level differences for each unit in turn.

3.1.2 Difference from PE mark split by examiner role

We calculated script level differences from the PE mark across all subjects in the study for examiners who occupied the TL role, and those who occupied an AE role. We were interested to know whether TLs marked more consistently than the AEs when compared to the PE. In advance we might predict that TLs would show less difference, since part of the selection for the team leader role includes an evaluation of the quality of their marking. Moreover, they are probably more familiar with the PE's way of marking and thinking from attending face-to-face pre-standardisation/standardisation meetings.

Averaged across every single unit in the study from both years, we found that in terms of their overall leniency/severity, TL script marks were no closer to the PE mark than those of AEs, with both marking 2.0% (as a proportion of whole paper mark) more leniently than the PE. In terms of their absolute difference from the PE mark (the size of their average difference, ignoring the direction), TLs were marginally closer to the PE, with a difference of 6.6% compared to AEs with a difference of 7.4% ($t(12588) = 5.80$, $p < 0.01$). TLs were therefore slightly less variable than AEs, clustering a little more closely around the PE's script mark than the AEs, as might be expected from their greater experience.

3.1.3 Difference from median mark

For each item, the median mark was obtained (including the PE mark) and script-level mark differences from this median, as a percentage of the marks available on the paper, were calculated for each script and marker combination (including the PE). The mean and standard deviations of these differences are shown for the 2016 units (Figure 4) and the 2017 units (Figure 5).

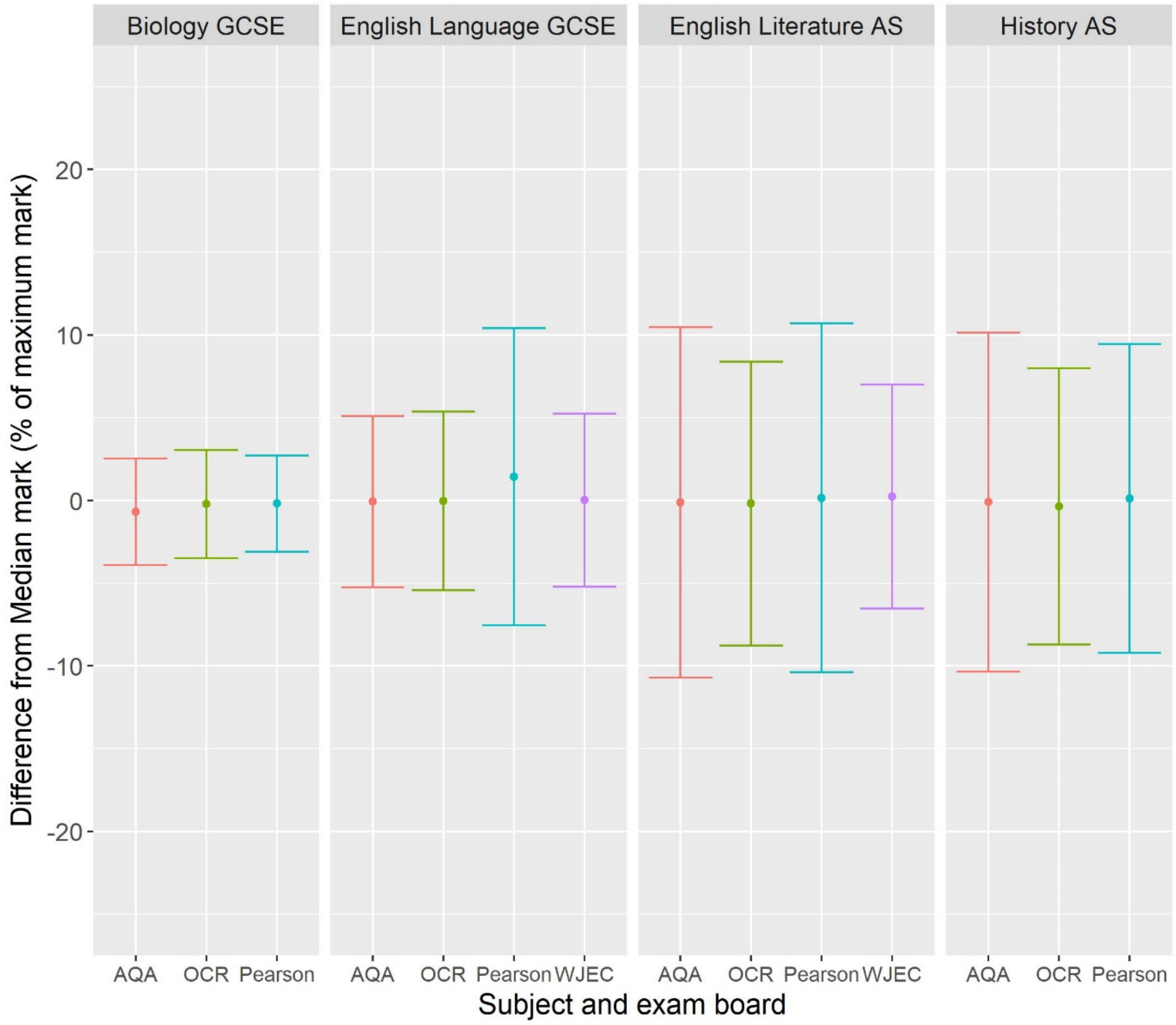


Figure 4: *Mean and standard deviation of whole script mark differences from the median mark, averaged across all markers and all scripts, for summer 2016 units.*

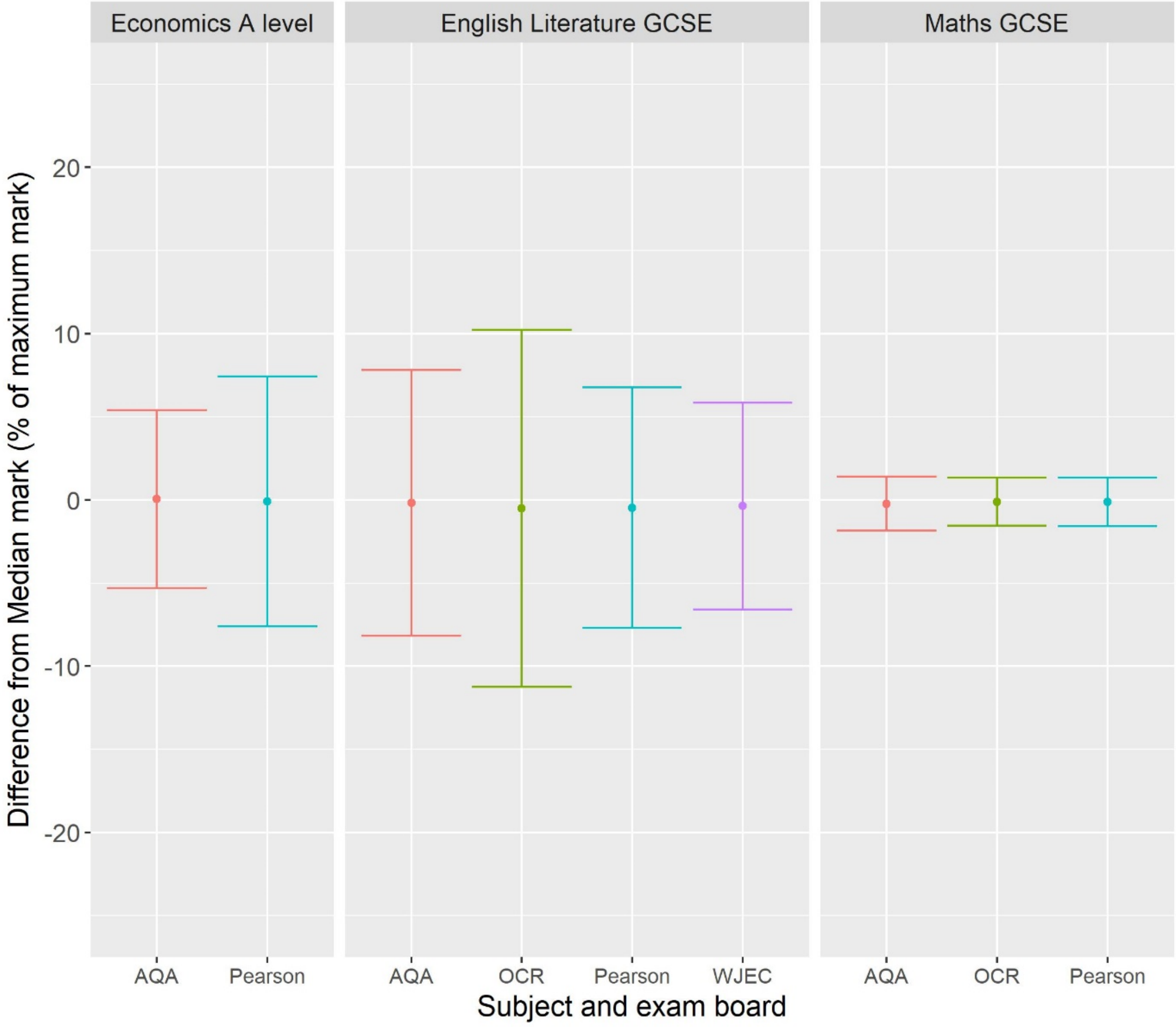


Figure 5: *Mean and standard deviation of whole script mark differences from the median mark, averaged across all markers and all scripts, for summer 2017 units.*

These script level differences from the median to a large extent reproduce the differences from the PE mark.

3.2 Item level variation

This section details mark differences between the markers and the PE mark, at the level of individual items. Corresponding data on marker differences to the median mark are given in Appendix D. Items are grouped by tariff, and the higher tariffs are grouped into tariff bands, in order to simplify the plots. In many cases where a paper contains a number of questions with the same (often high) tariffs they are optional questions, and so treating them as equivalent rather than plotting individual questions is more representative of general accuracy. All of the figures in this section plot mark difference in actual marks, not scaled as a proportion of the maximum mark.

Individual plots for the 2016 units are shown in Figure 6 to Figure 9, while those for the 2017 units are shown in Figure 10 to Figure 12. Not every exam board appears in every tariff band due to the pattern of tariffs on each paper.

2016 units

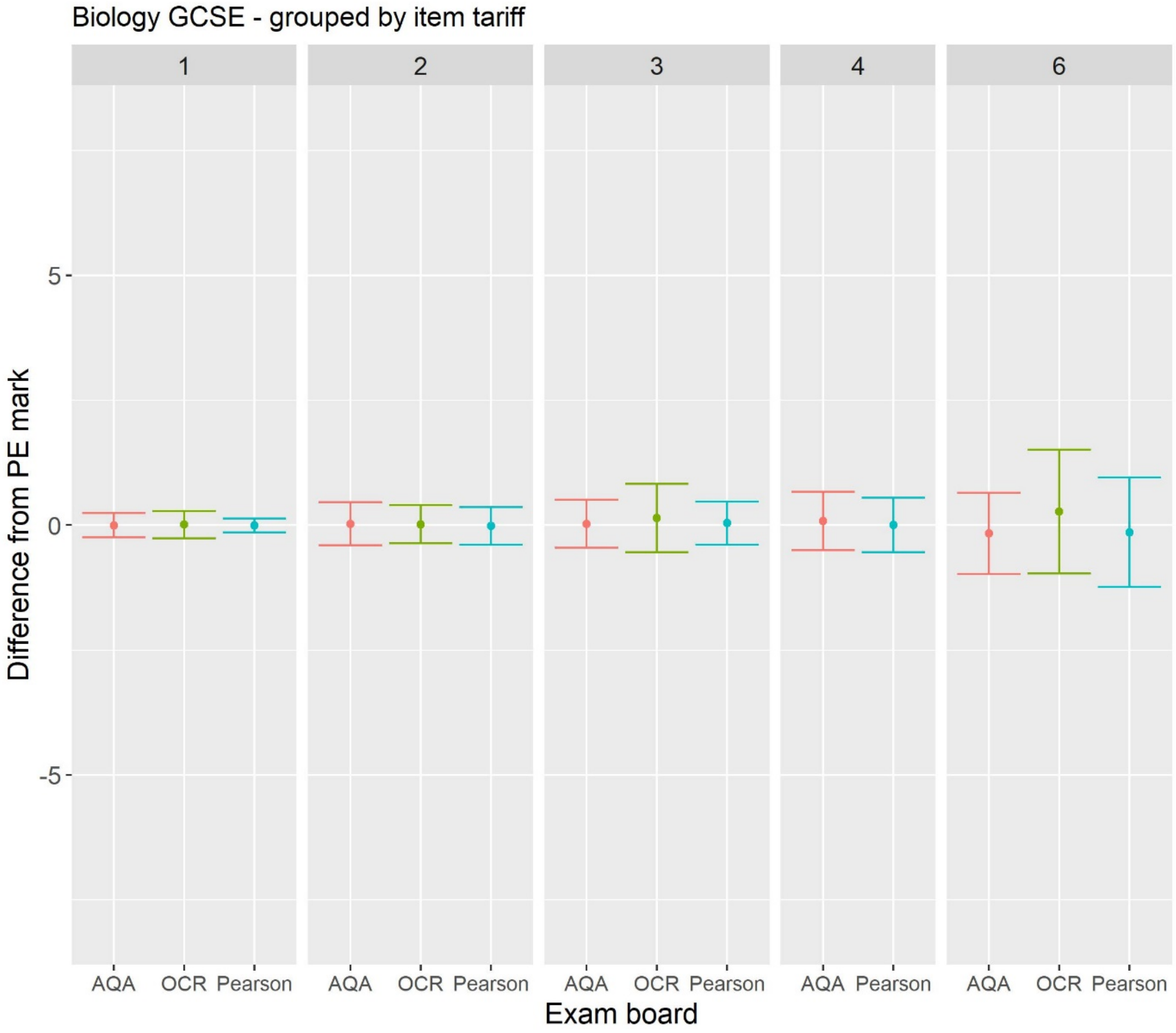


Figure 6: *Mean and standard deviation of the difference of item marks from the PE mark, averaged across all markers and grouped by item tariff, for the summer 2016 biology GCSE units. The exam boards are plotted separately and in different colours.*

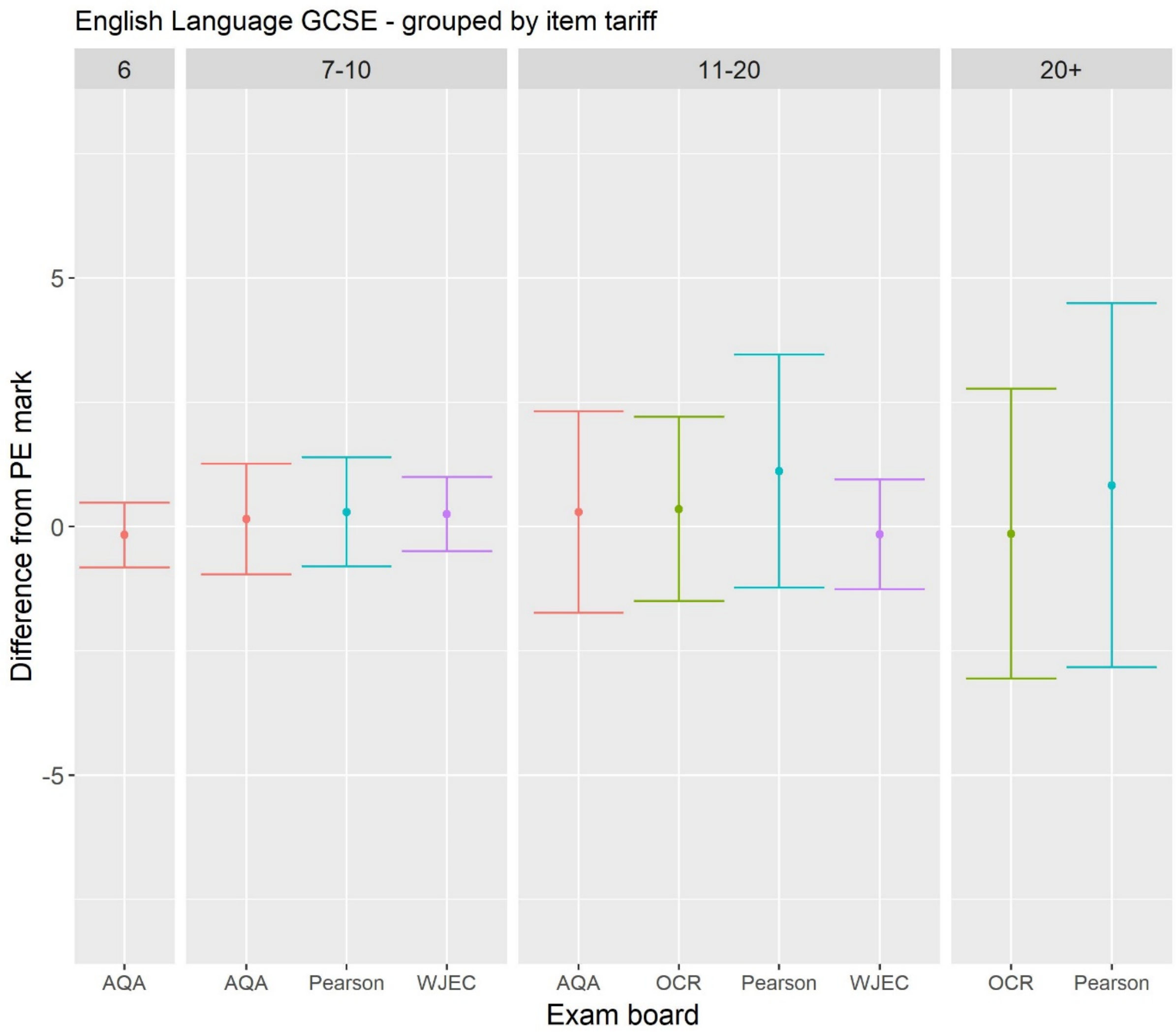


Figure 7: *Mean and standard deviation of the difference of item marks from the PE mark, averaged across all markers and grouped by item tariff, for the summer 2016 English language GCSE units. The exam boards are plotted separately and in different colours.*

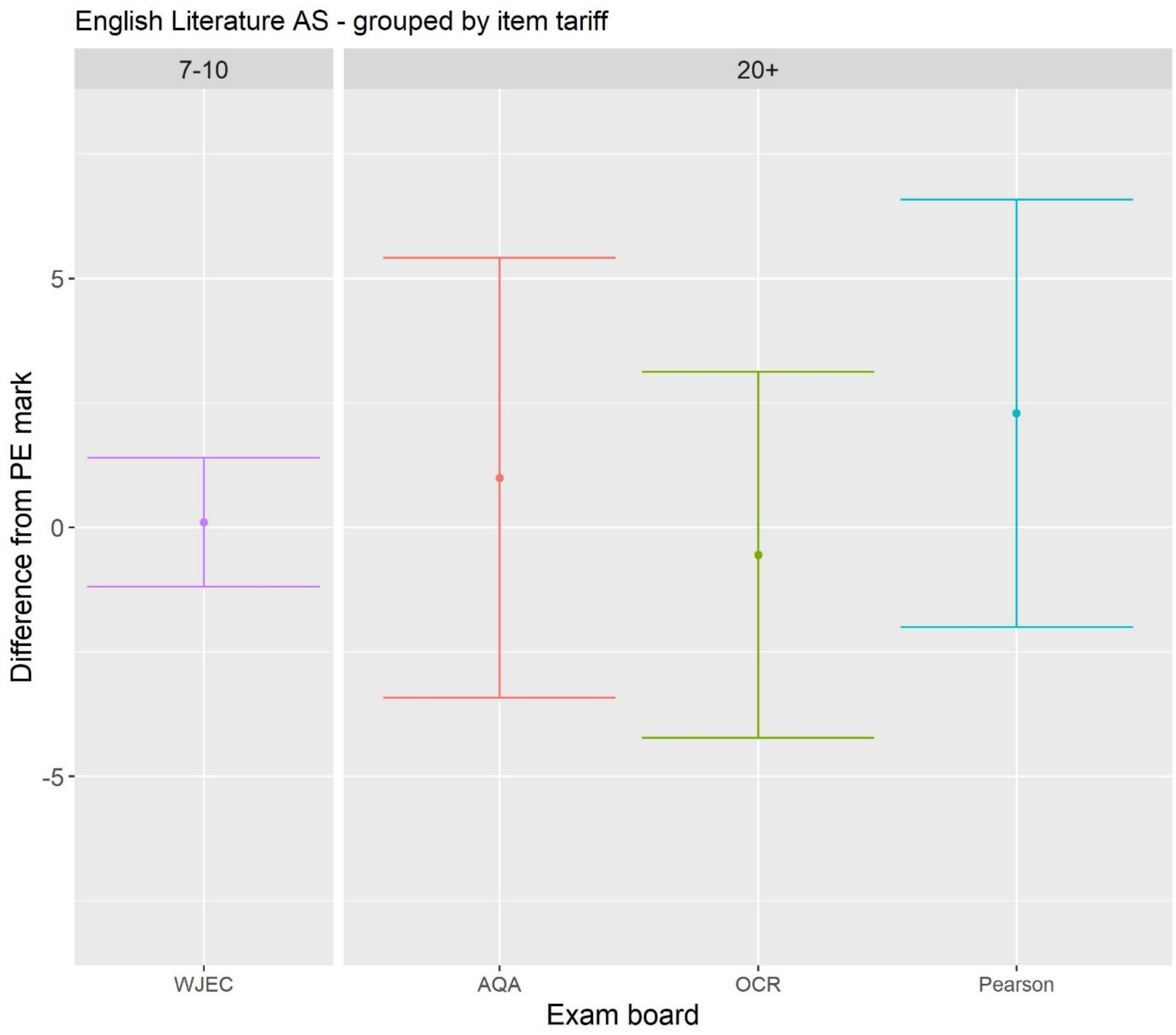


Figure 8: *Mean and standard deviation of the difference of item marks from the PE mark, averaged across all markers and grouped by item tariff, for the summer 2016 English literature AS units. The exam boards are plotted separately and in different colours.*

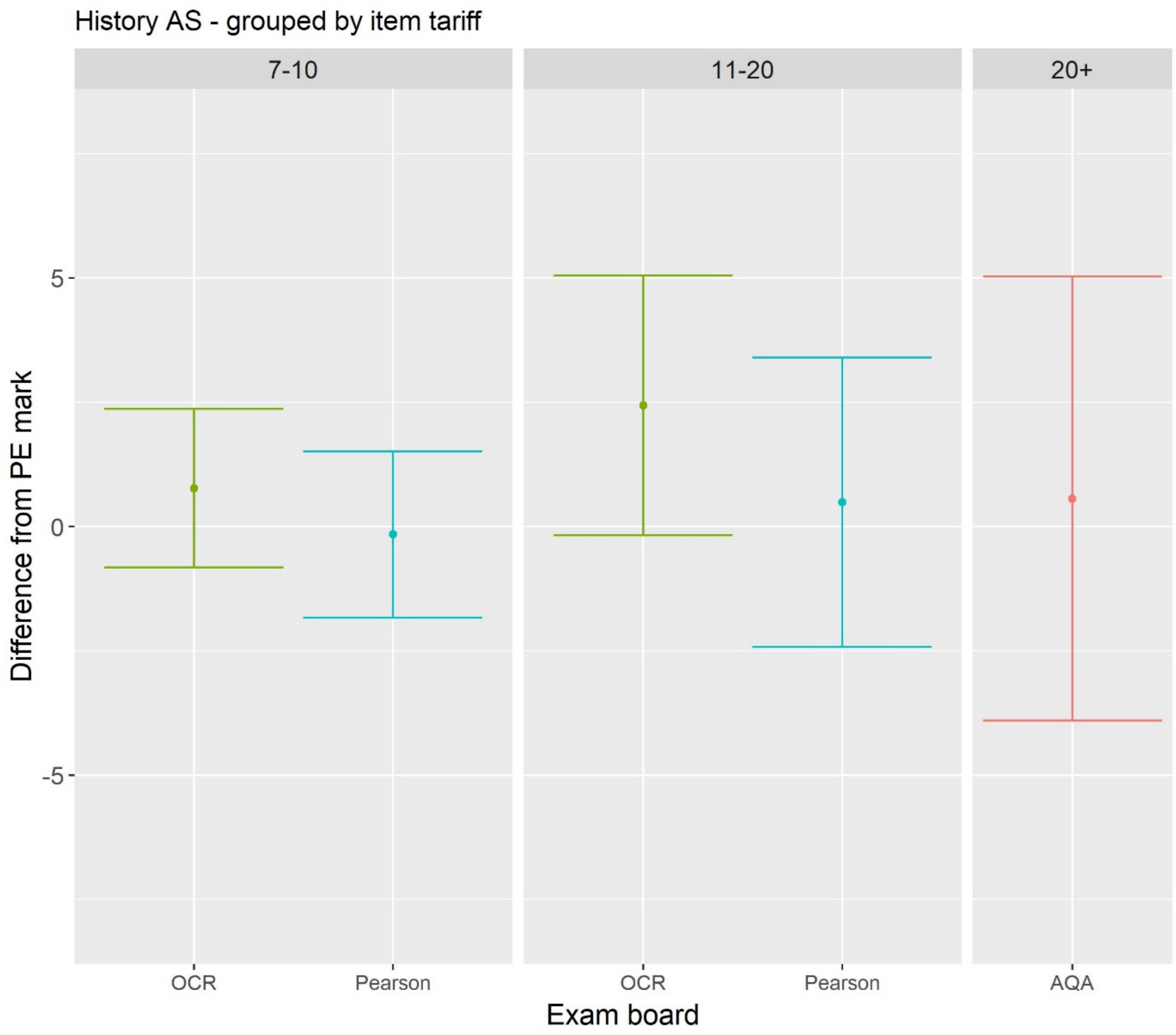


Figure 9: *Mean and standard deviation of the difference of item marks from the PE mark, averaged across all markers and grouped by item tariff, for the summer 2016 history AS units. The exam boards are plotted separately and in different colours.*

2017 units

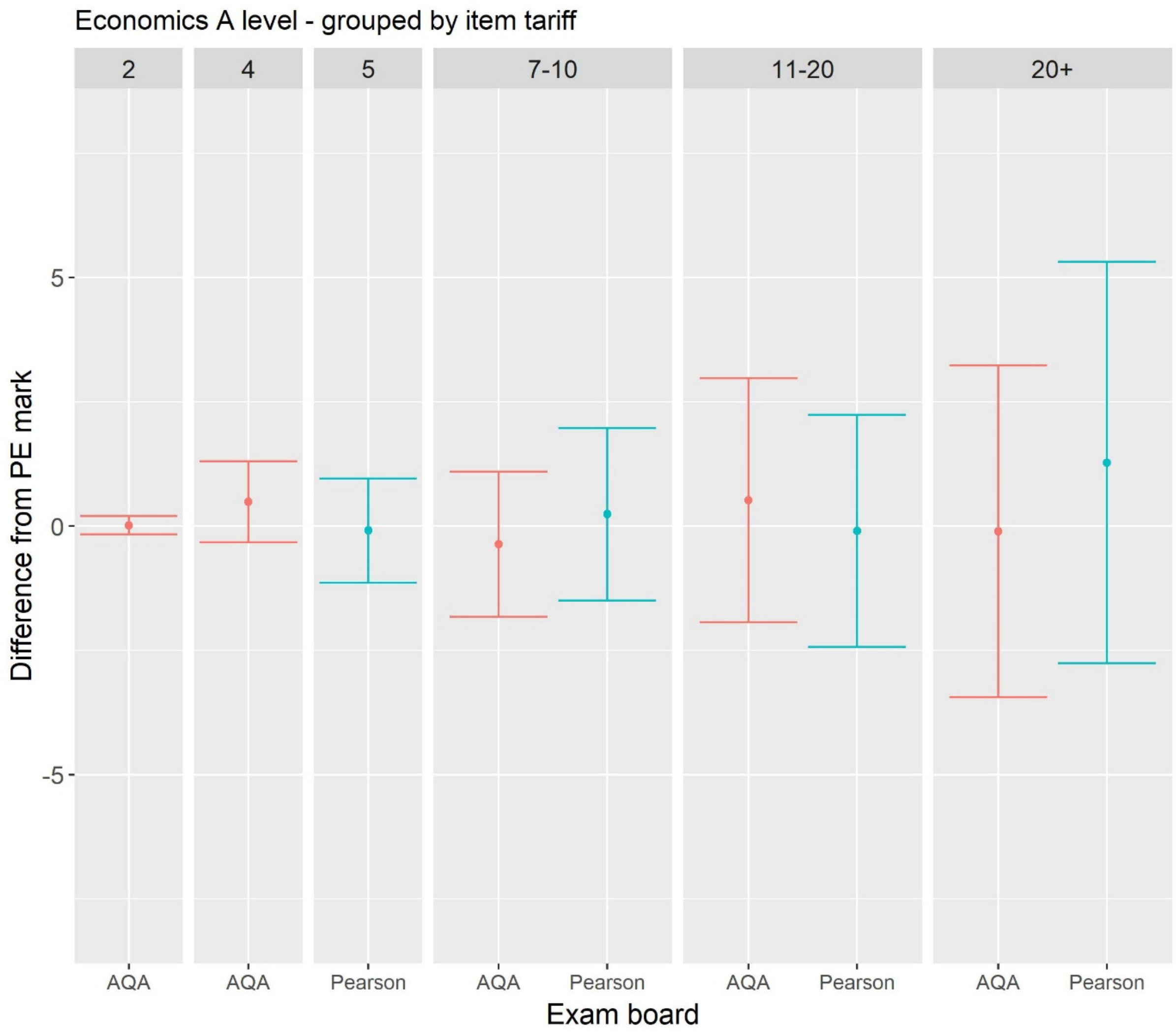


Figure 10: *Mean and standard deviation of the difference of item marks from the PE mark, averaged across all markers and grouped by item tariff, for the summer 2017 economics A level units. The exam boards are plotted separately and in different colours.*

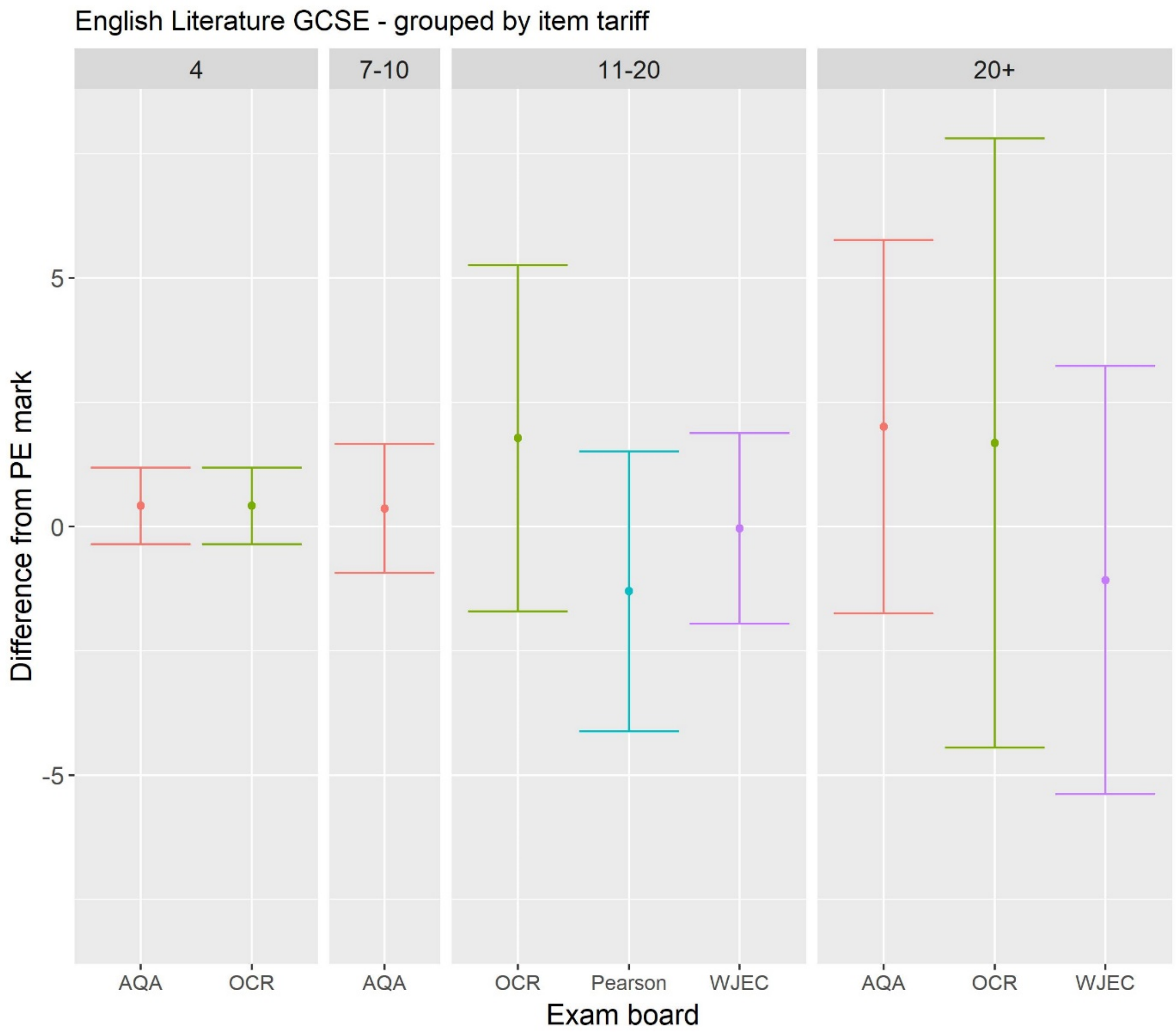


Figure 11: *Mean and standard deviation of the difference of item marks from the PE mark, averaged across all markers and grouped by item tariff, for the summer 2017 English literature GCSE units. The exam boards are plotted separately and in different colours.*

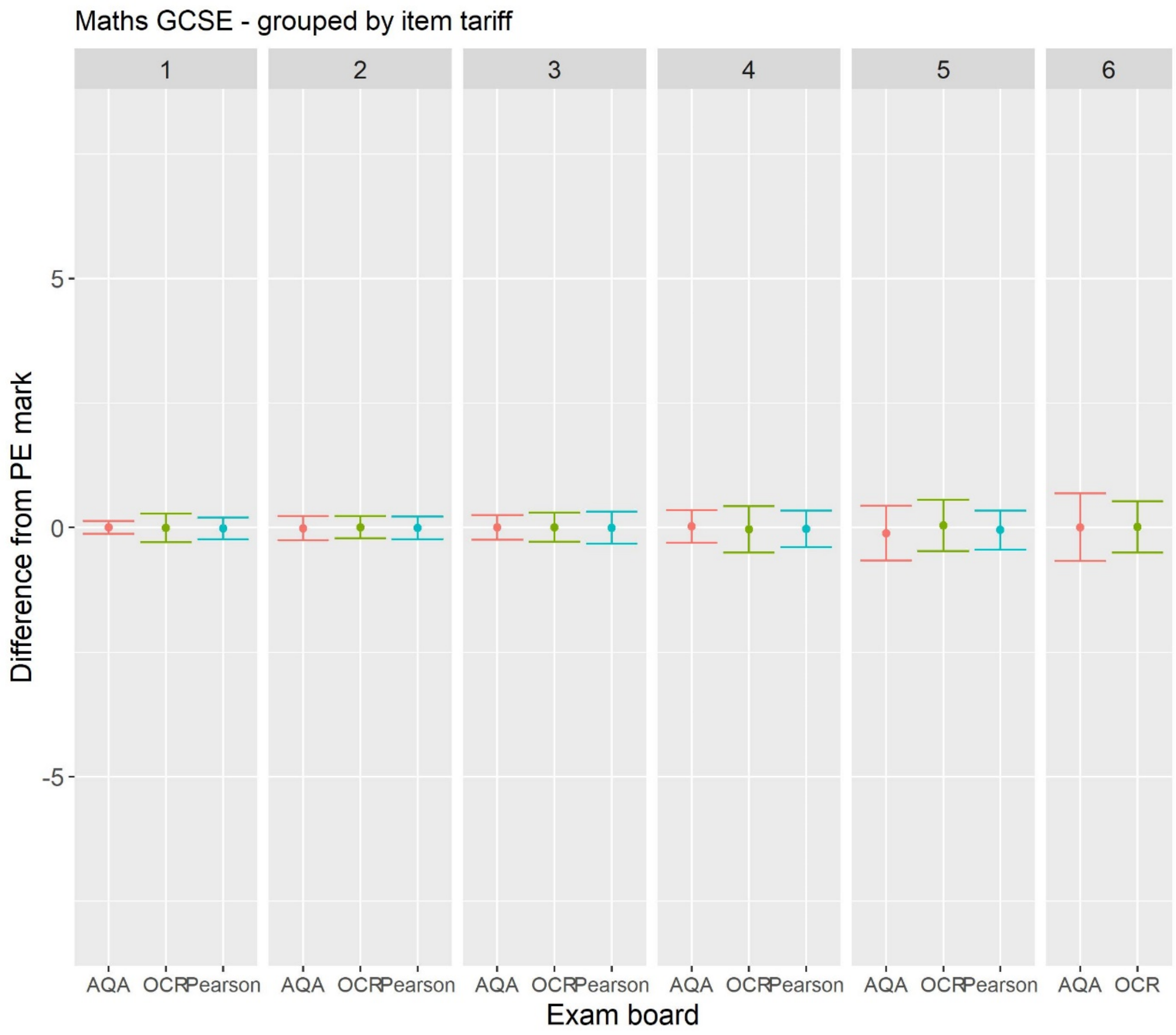


Figure 12: *Mean and standard deviation of the difference of item marks from the PE mark, averaged across all markers and grouped by item tariff, for the summer 2017 mathematics GCSE units. The exam boards are plotted separately and in different colours.*

Apart from the differences between units within each tariff band, the spread of mark differences is larger for the higher tariff questions than the lower tariff ones, simply due to the greater number of marks available over which disagreement could occur. As a general rule, the size of the standard deviation is proportional to the mark of the item.

3.3 Survey responses

At the end of the marking exercise, we asked all of our examiners to complete a short online survey. We asked questions about the marking itself, in terms of how similar to the live marking it was, and also asked about the use of the marking software and scope for improvements in order to guide refinements to the software. Here we only report on questions relating to the marking, split by year, since the marking software differed slightly and this may have influenced some respondents (see Figure 13).

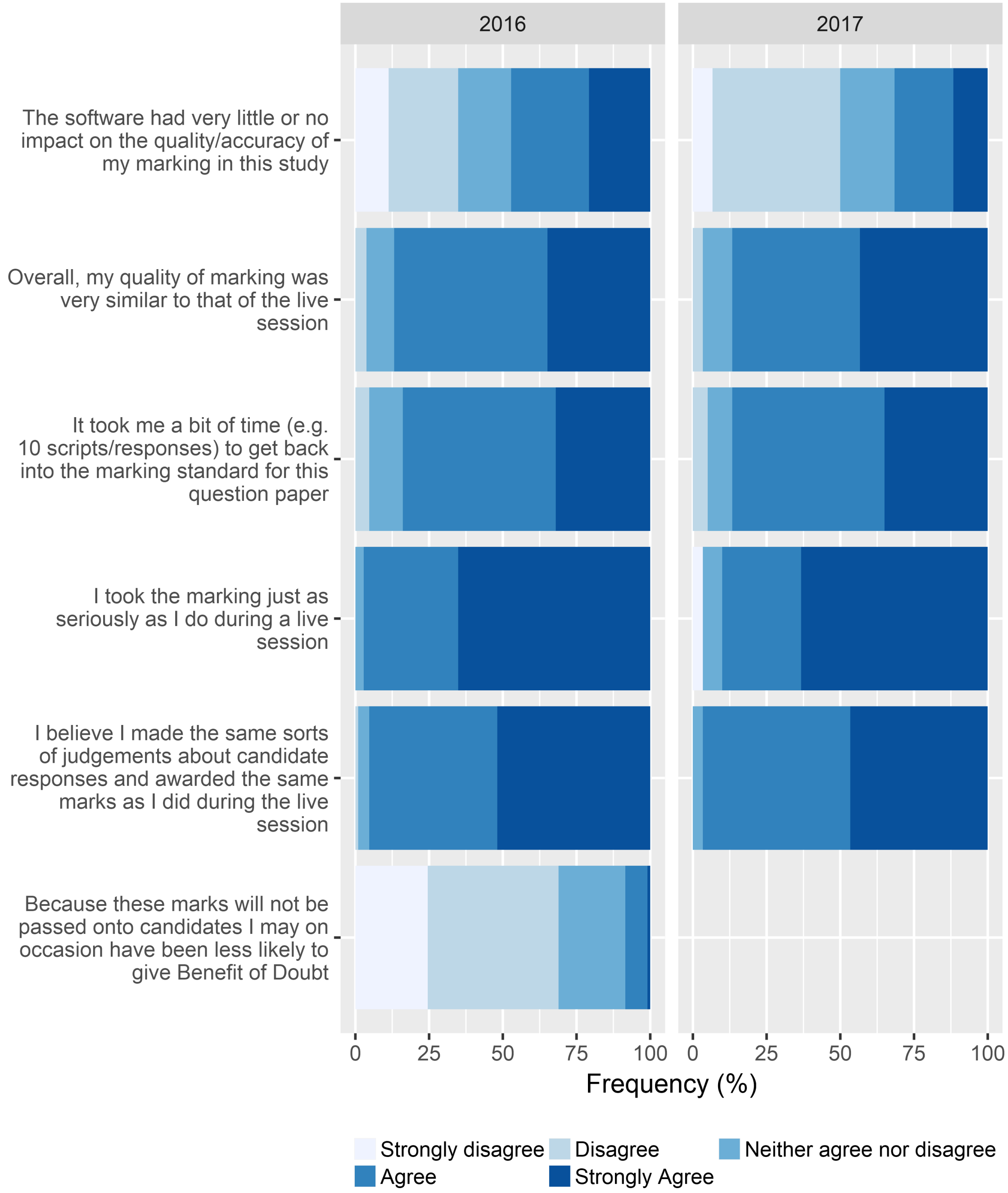


Figure 13: *Analysis of the Likert-scale questions from the examiner survey. The responses are combined across units but split between the examiners who took part in the 2016 (left column) and 2017 (right column) studies. The final question was not asked in 2017.*

Clearly, although the marking software might have had some impact, the majority of examiners took this task very seriously and felt that they had marked very similarly to how they marked in the live summer marking. One open response question asked for any other comments on the marking. While many examiners raised no specific issues and stated that the marking should be very similar and they had applied the same standard, there were a few concerns. A number of comments were made about the lack of annotations in the software for the 2016 units (as well as some limitations with the annotations for 2017 units) and the effect the consequent inability to note down thought processes may have had on the quality or consistency of the marking. In some units with optional questions, examiners sometimes noted that in the live marking they had marked sets of responses to specific questions, while in this study the sample was random. This meant that examiners had to cope with a wider range of questions being answered within a section of a paper, which may have impacted on accuracy.

Comments were also made about the difference of the current process from the closely monitored and managed live marking, with no senior examiner to speak to about problems and no seeds to keep examiners on-standard. Some examiners did comment positively about the software, including the flexibility of the system in how they marked and that it kept all marked scripts/items accessible to the end, allowing examiners to adjust marks upon further reflection, perhaps leading to greater consistency.

Overall, it does appear that allowing for some differences in the process and the software, most examiners applied themselves well to this task and marked as similarly to the summer session as they were able to, despite the lower-stakes nature of this marking exercise compared to live summer marking.

4 Discussion

The data presented here agree very closely with measures obtained from the exam boards' own monitoring data (the marking consistency metrics), although the differences we have measured tend to be slightly larger than in the metrics. Two factors may contribute to the slightly larger disagreement. First, we asked for an entirely random selection of scripts, so that the sample may have contained scripts which were more difficult to mark than scripts/items used in normal live marker monitoring. Very difficult to mark responses rarely get chosen as seed items, since seeds should be fairly unambiguous when applying the mark scheme to them. Therefore on average, the responses in the current sample are likely to be slightly harder to mark than the seeds that form the basis of the data in the marking consistency metrics. Second, although the survey responses suggest that markers were taking this exercise seriously and marked as they would have done in the summer series, this was a much less pressurised situation, and we cannot rule out that individuals may not have thought quite so long and hard about a final mark, since they knew this marking had no consequences for any candidates and they were not at risk of being stopped from marking if they did not mark close to the PE standard. These factors, together with the relative unfamiliarity of the marking software, could have led to slightly more varied marking. This does not invalidate the relative differences between units though, as these factors would have been relatively constant across units. Differences between units will therefore represent genuine differences in the consistency of marking. It is worth noting here that paper and question structure can influence marking consistency. Papers from different exam boards can have different distributions of item tariffs, and some boards break question marks up into the constituent assessment objectives (where more than one is being assessed) in different ways. We did not analyse these factors in the units we studied, but note their potential influence on the data we collected.

Finally, we cannot be sure which of the 2 sets of data, marking consistency metrics or the marking consistency data presented here, are ultimately more representative of the differences that would occur across all scripts/items in live marking. The marking consistency metrics data, being based on seed items, may underestimate the size of disagreements due to the less ambiguous nature of the seeds relative to the whole set of candidate responses to an item. The current marking consistency data probably slightly overestimates the variability, due to the reasons given above.

5 Conclusions

This marking consistency study showed that there are differences in the size of marker disagreement between units, both within and between subjects. The spread of marker disagreement with the definitive (PE) mark corresponds very closely with the same measures from the marking consistency metrics. Therefore we have confidence in both approaches as measures of the underlying consistency of marking at unit/component level.

Appendix A – Simulations of the effect of distributed item marking on the standard deviation of whole script mark differences

For many online-marked papers, items from a single candidate script are distributed across multiple markers in a random allocation. In the current study, whole scripts were loaded onto the marking system, so that even when examiners chose to mark by item, working through all responses to each question in turn, they were still marking every item on every script. Our measures of whole script difference are therefore based on individual examiner's script totals. One potential advantage of distributed item marking is that any systematic marking bias (severity or leniency) by an individual marker is reduced when they mark few items on any one script. Therefore our whole-script marking may have inflated script differences when compared to the distributed item marking in the live session.

In order to determine the extent of any difference, we carried out simulations of whole script mark differences, in which script totals were constructed by randomly sampling marks for each item from all the examiners in the study. By repeating this random mark selection 50 times, we obtain an estimate of the expected variation of script mark differences arising with distributed item marking, which can be compared to our whole script differences.

We found that across the whole set of units, distributed item marking led to a 10.1% reduction in the standard deviation of whole script mark differences. Table A1 details the effect observed for each subject.

Table A1: *Change in the standard deviation of whole script mark differences when switching from whole script marking to distributed item marking. Percentage change is averaged for each subject across the units for that subject.*

Subject	% change in standard deviation of the mark difference for distributed marking relative to whole script marking
Biology GCSE	-4.6%
English Language GCSE	-15.1%
English Literature AS	-12.9%
History AS	-9.4%
Economics A level	-10.6%
English Literature GCSE	-12.8%
Mathematics GCSE	-2.3%

The measures of whole script mark deviations we obtained in our study are slightly larger than might be expected from distributed marking, and it is larger for units with predominantly higher-tariff questions and those associated with less reliable marking. We have not made any correction to our data given that, although it is unlikely, examiners were free to mark either by script or by item in our study. The magnitude of the effect may also be slightly different given the different circumstances of our study.

Appendix B – Whole script mark difference histograms

This appendix contains histograms of whole script mark differences from the PE mark for all units in the study, plotted in marks (Figures B1 to B7). Since the papers differ by maximum mark this will affect the spread of mark differences. All figures use the same scale along the bottom (x-) axis in order to allow direct comparison.

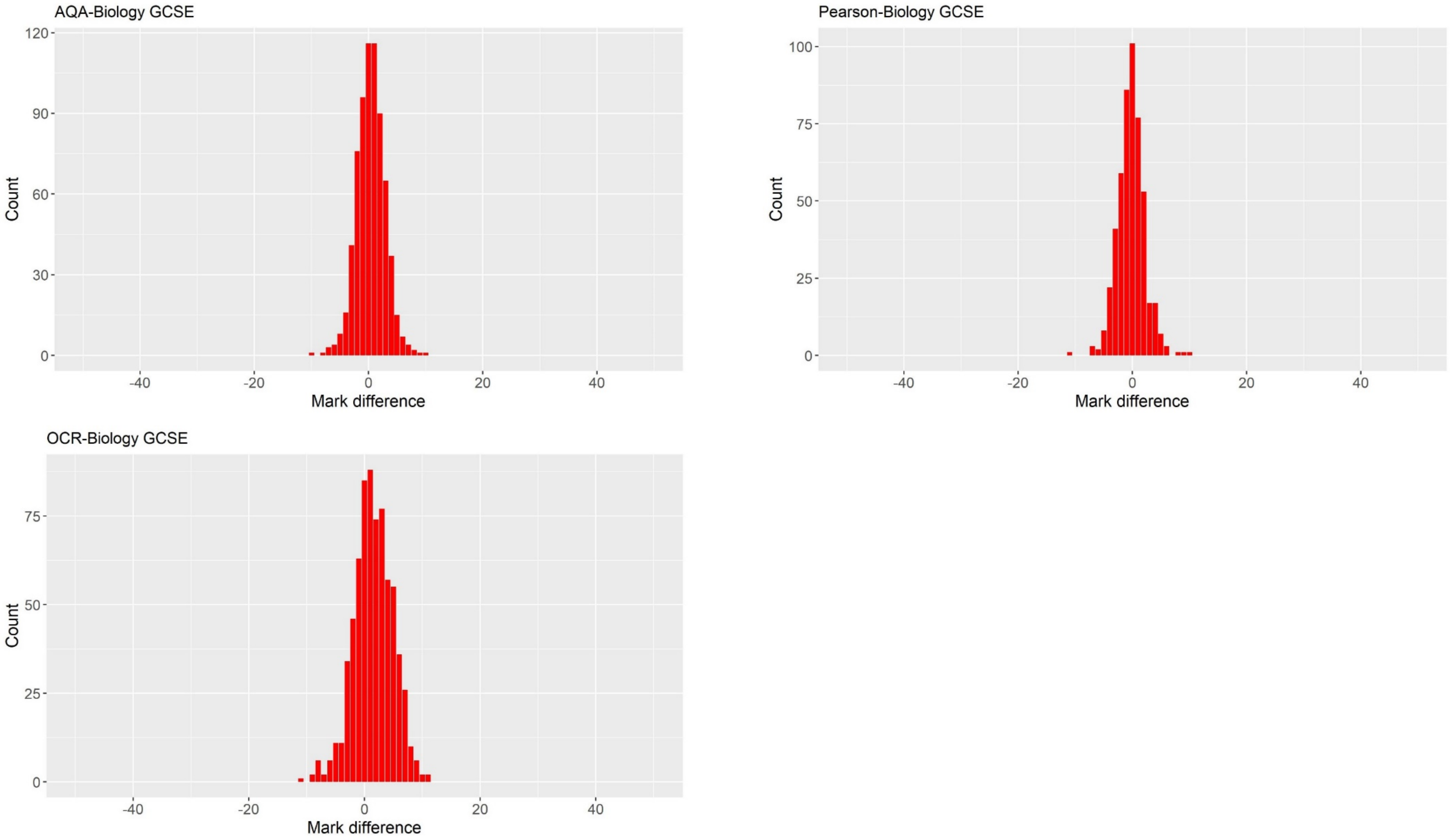


Figure B1: *Distribution of whole script mark differences from the PE mark across all markers and all scripts, for 2016 biology GCSE.*

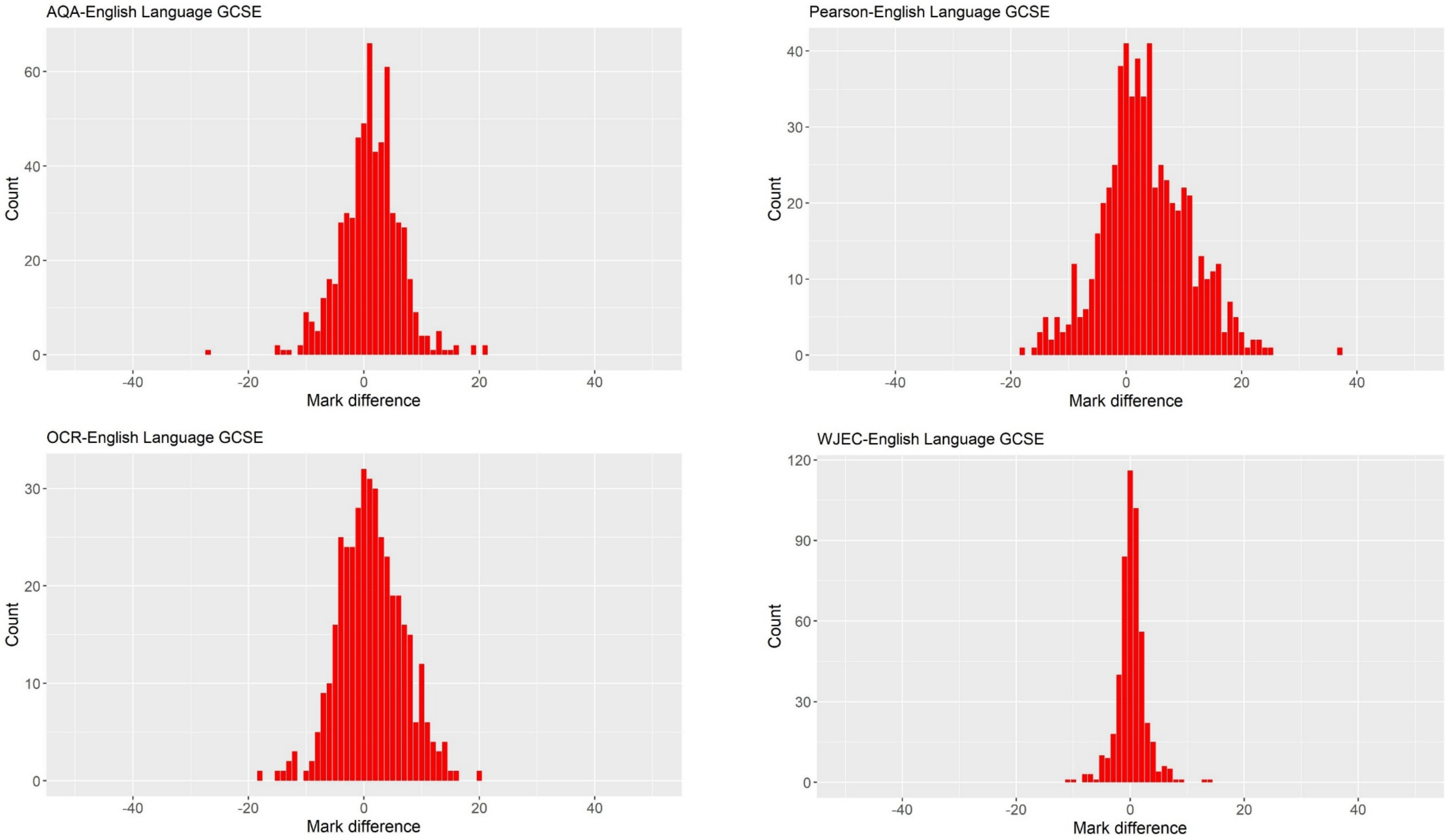


Figure B2: *Distribution of whole script mark differences from the PE mark across all markers and all scripts, for 2016 English language GCSE*.

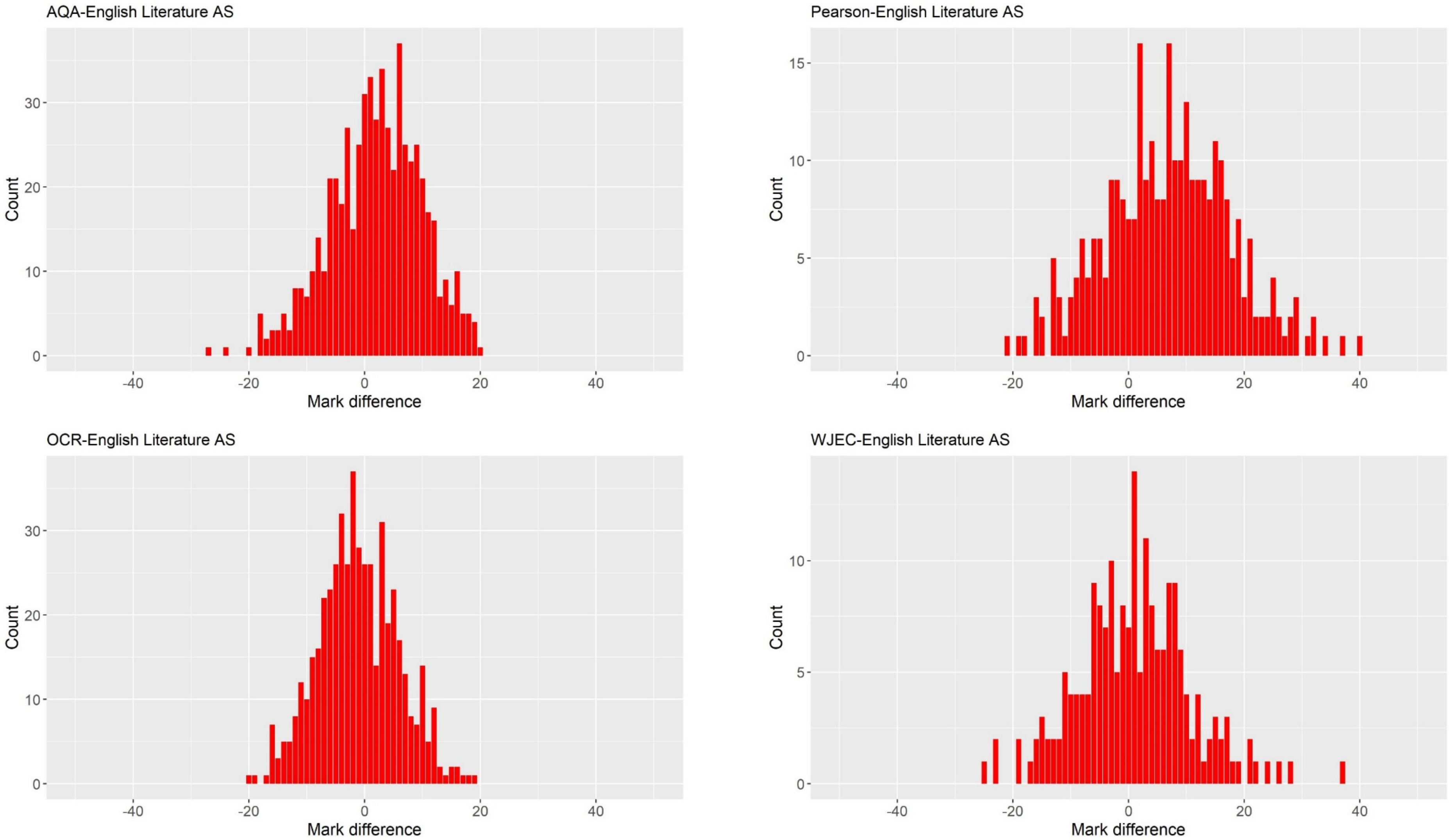


Figure B3: *Distribution of whole script mark differences from the PE mark across all markers and all scripts, for 2016 English literature AS.*

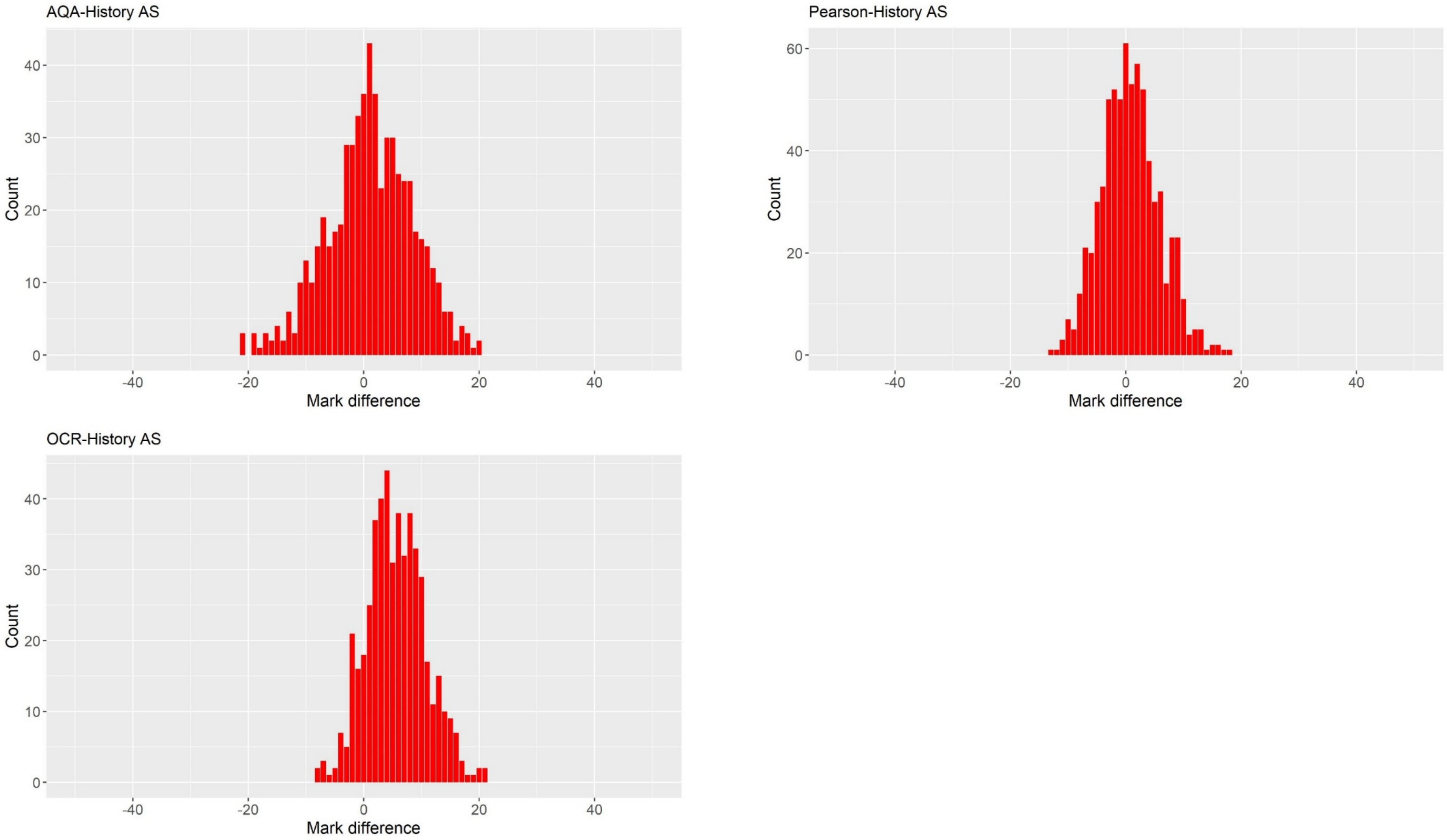


Figure B4: *Distribution of whole script mark differences from the PE mark across all markers and all scripts, for 2016 history AS.*

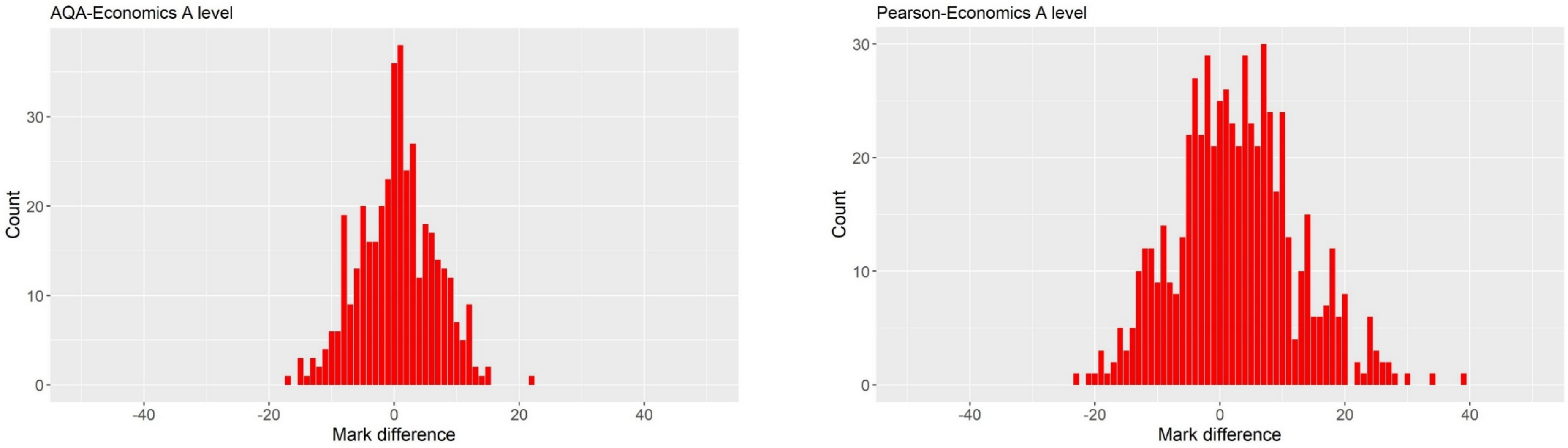
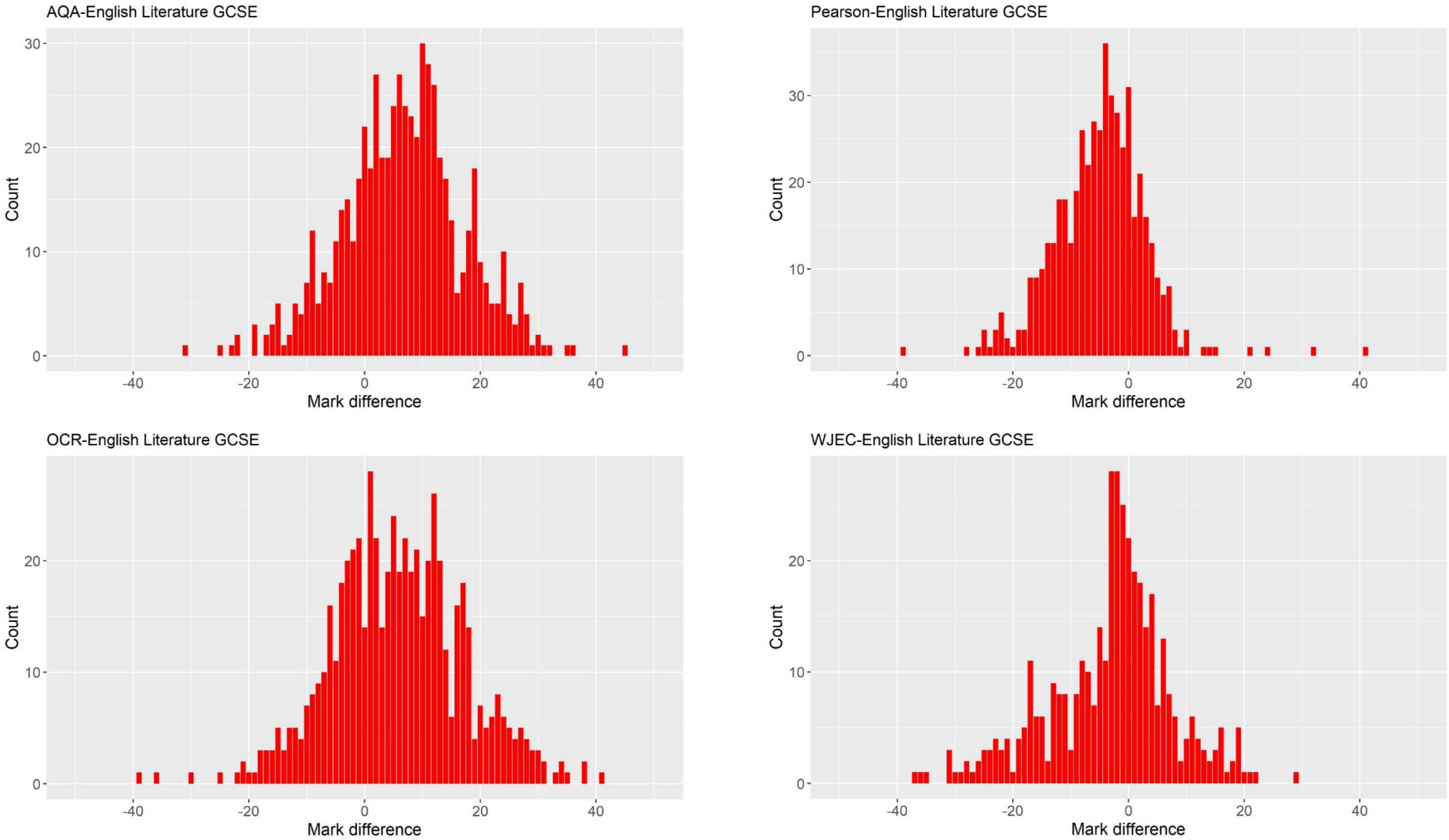


Figure B5: *Distribution of whole script mark differences from the PE mark across all markers and all scripts, for 2017 economics A level.*



AQA-English Literature GCSE

Pearson-English Literature GCSE

OCR-English Literature GCSE

WJEC-English Literature GCSE

Figure B6: *Distribution of whole script mark differences from the PE mark across all markers and all scripts, for 2017 English literature GCSE.*

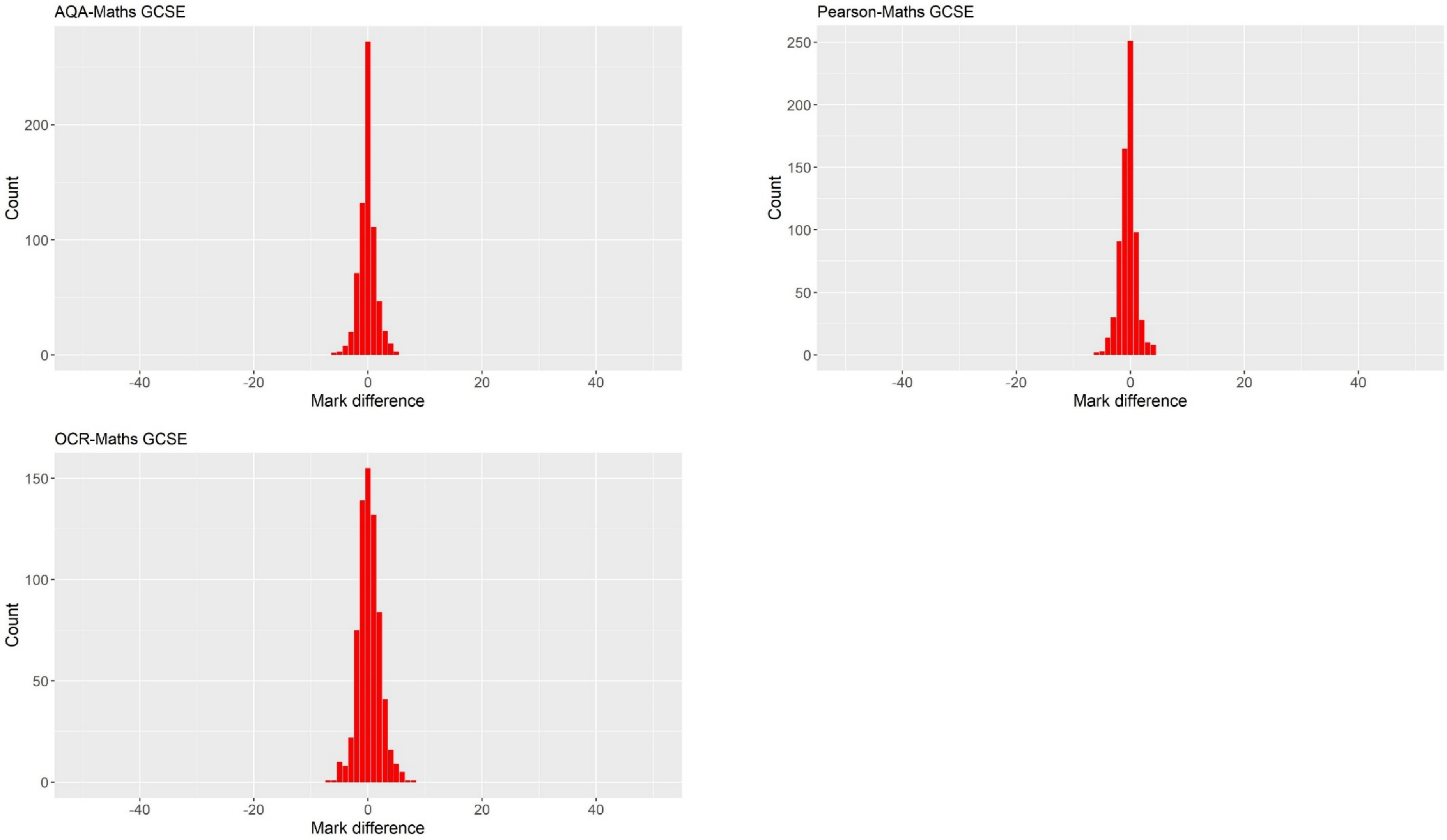


Figure B7: *Distribution of whole script mark differences from the PE mark across all markers and all scripts, for 2017 mathematics GCSE.*

Appendix C – Additional data tables

Table C1: *Mean and standard deviation of whole script mark differences from the PE mark, averaged across all markers and all scripts, for summer 2016 units.*

Subject	Mean (% of whole script mark)	Standard deviation (% of whole script mark)
Biology GCSE		
AQA	0.67	4.16
OCR	2.02	4.48
Pearson	-0.39	3.95
English language GCSE		
AQA	1.44	6.47
OCR	1.58	6.91
Pearson	5.24	11.73
WJEC	0.50	6.34
English literature AS		
AQA	3.97	15.82
OCR	-1.84	11.41
Pearson	9.52	15.14
WJEC	1.01	9.78
History AS		
AQA	2.24	14.79
OCR	11.24	10.19
Pearson	2.05	12.69

Table C2: *Mean and standard deviation of whole script mark differences from the PE mark, averaged across all markers and all scripts, for summer 2017 units.*

Subject	Mean (% of whole script mark)	Standard deviation (% of whole script mark)
Economics A level		
AQA	0.55	7.60
Pearson	2.63	9.64
English literature GCSE		
AQA	6.88	10.88
OCR	7.04	14.27
Pearson	-6.25	9.90
WJEC	-2.64	8.90
Maths GCSE		
AQA	-0.13	1.88
OCR	0.18	1.96
Pearson	-0.54	1.81

Table C3: *Mean and standard deviation of whole script mark differences from the median mark, averaged across all markers and all scripts, for summer 2016 units.*

Subject	Mean (% of whole script mark)	Standard deviation (% of whole script mark)
Biology GCSE		
AQA	-0.68	3.23
OCR	-0.21	3.26
Pearson	-0.18	2.90
English language GCSE		
AQA	-0.07	5.16
OCR	-0.03	5.40
Pearson	1.44	8.97
WJEC	0.03	5.24
English literature AS		
AQA	-0.11	10.58
OCR	-0.18	8.58
Pearson	0.16	10.53
WJEC	0.23	6.77
History AS		
AQA	-0.10	10.23
OCR	-0.36	8.35
Pearson	0.13	9.33

Table C4: *Mean and standard deviation of whole script mark differences from the median mark, averaged across all markers and all scripts, for summer 2017 units.*

Subject	Mean (% of whole script mark)	Standard deviation (% of whole script mark)
Economics A level		
AQA	0.05	5.35
Pearson	-0.09	7.52
English literature GCSE		
AQA	-0.17	7.99
OCR	-0.50	10.73
Pearson	-0.46	7.23
WJEC	-0.37	6.22
Maths GCSE		
AQA	-0.22	1.62
OCR	-0.11	1.44
Pearson	-0.11	1.46

Appendix D – Item–level difference from median mark

Individual plots for the 2016 units are shown in Figures D1 to D4, while those for the 2017 units are shown in Figures D5 to D7. Not every exam board appears in every tariff band due to the pattern of tariffs on each paper.

2016 units

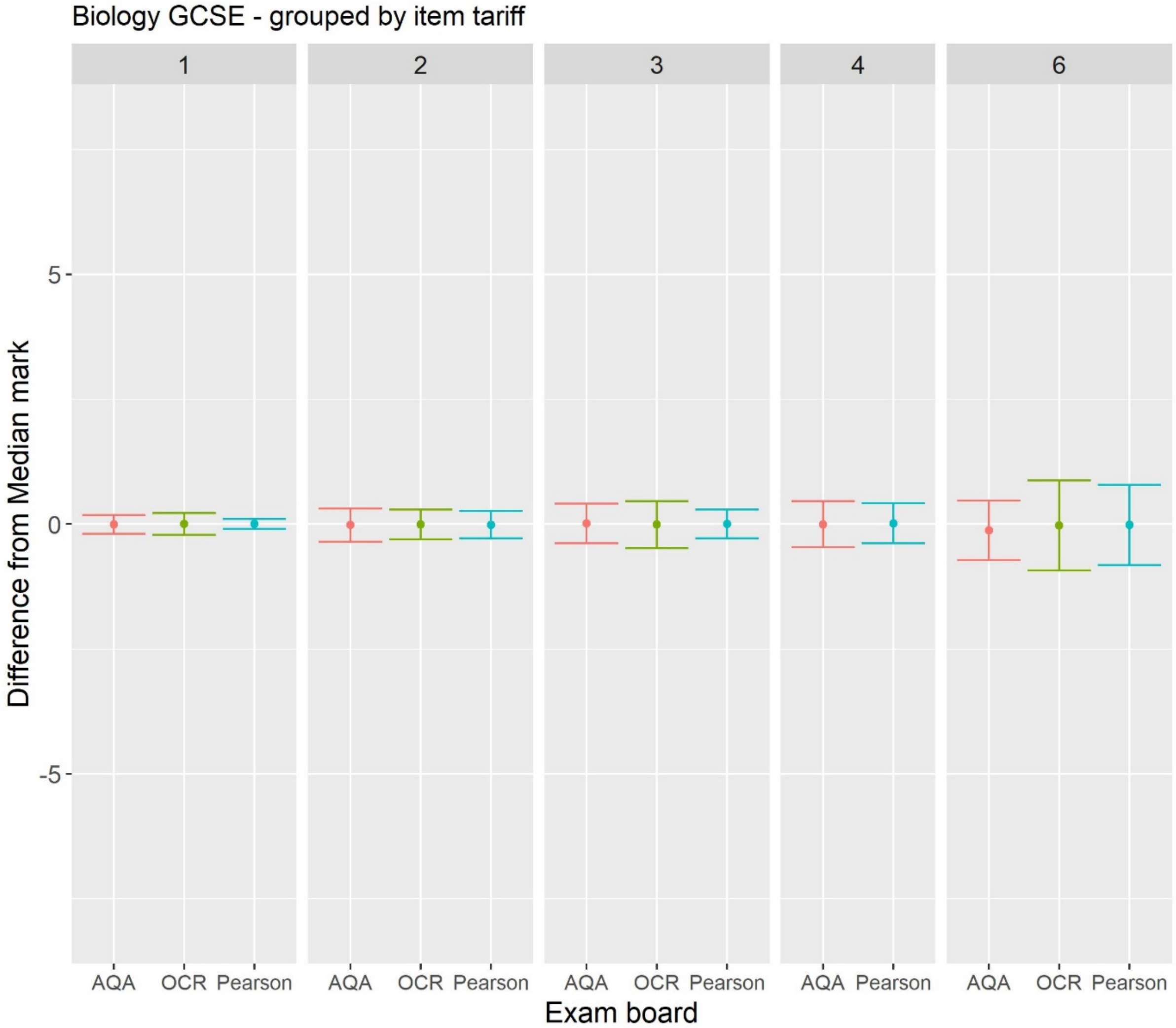


Figure D1: *Mean and standard deviation of the difference of item marks from the median mark, averaged across all markers and grouped by item tariff, for the summer 2016 biology GCSE units. The exam boards are plotted separately and in different colours*.

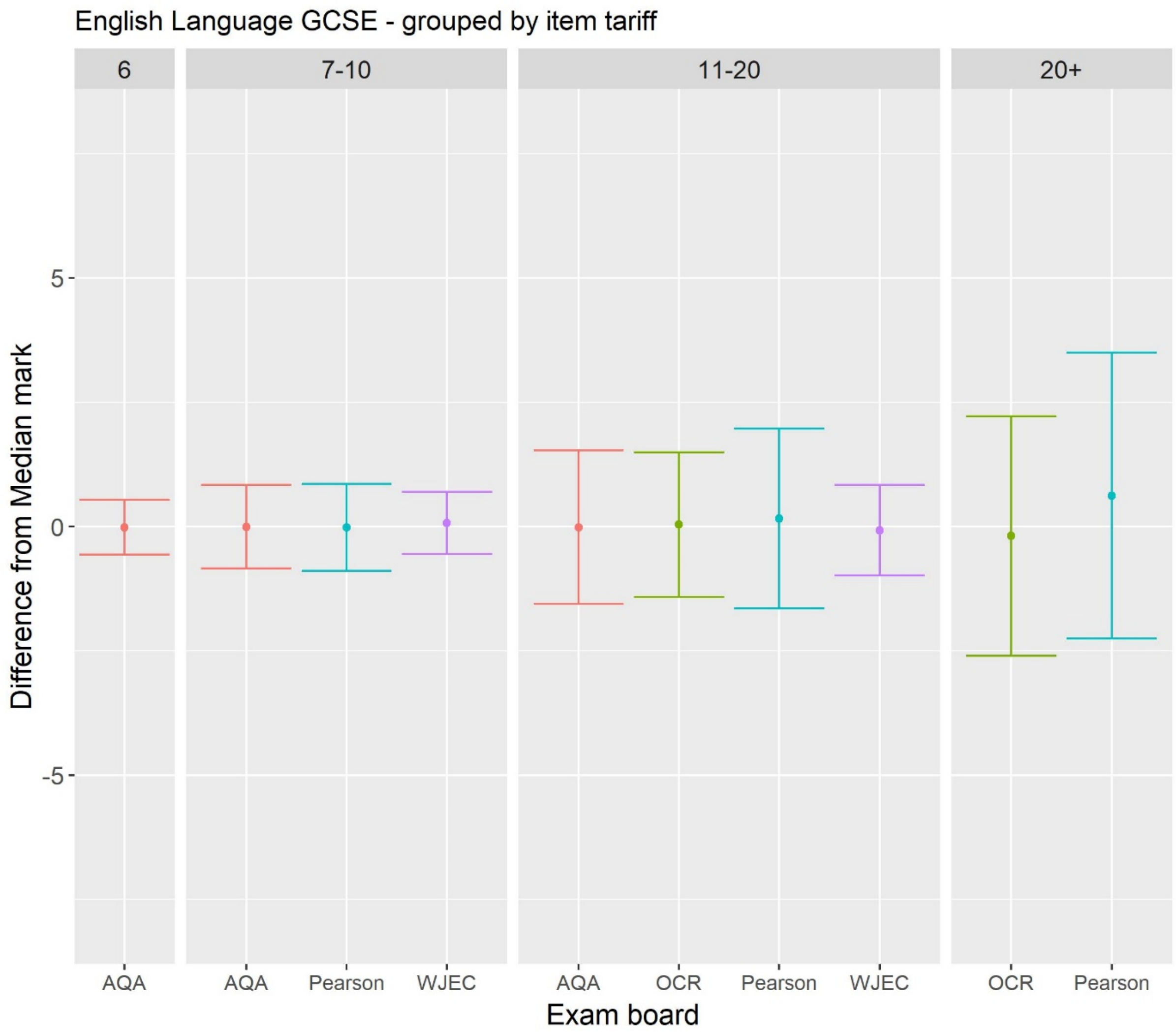


Figure D2: *Mean and standard deviation of the difference of item marks from the median mark, averaged across all markers and grouped by item tariff, for the summer 2016 English language GCSE units. The exam boards are plotted separately and in different colours.*

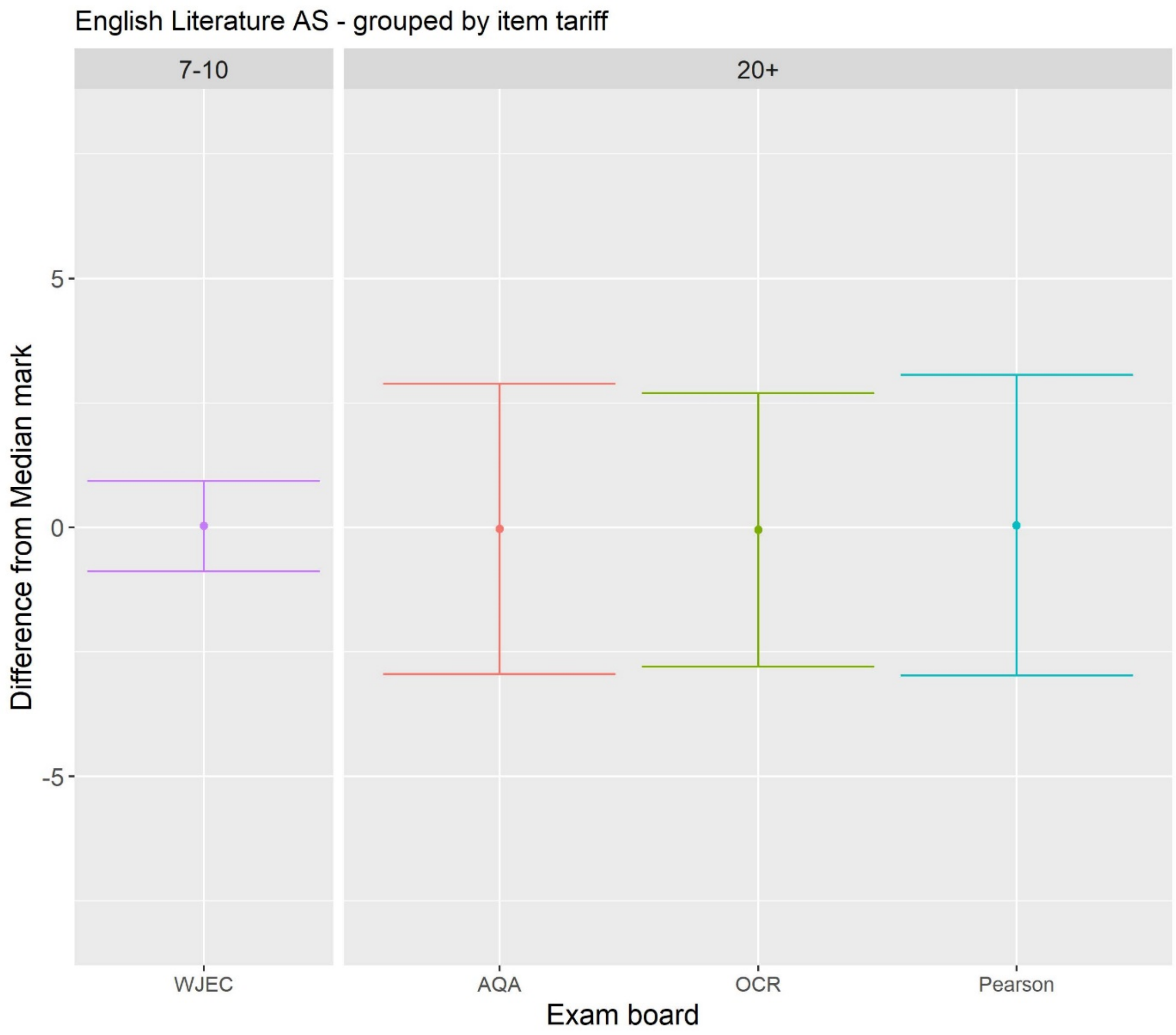


Figure D3: *Mean and standard deviation of the difference of item marks from the median mark, averaged across all markers and grouped by item tariff, for the summer 2016 English literature AS units. The exam boards are plotted separately and in different colours.*

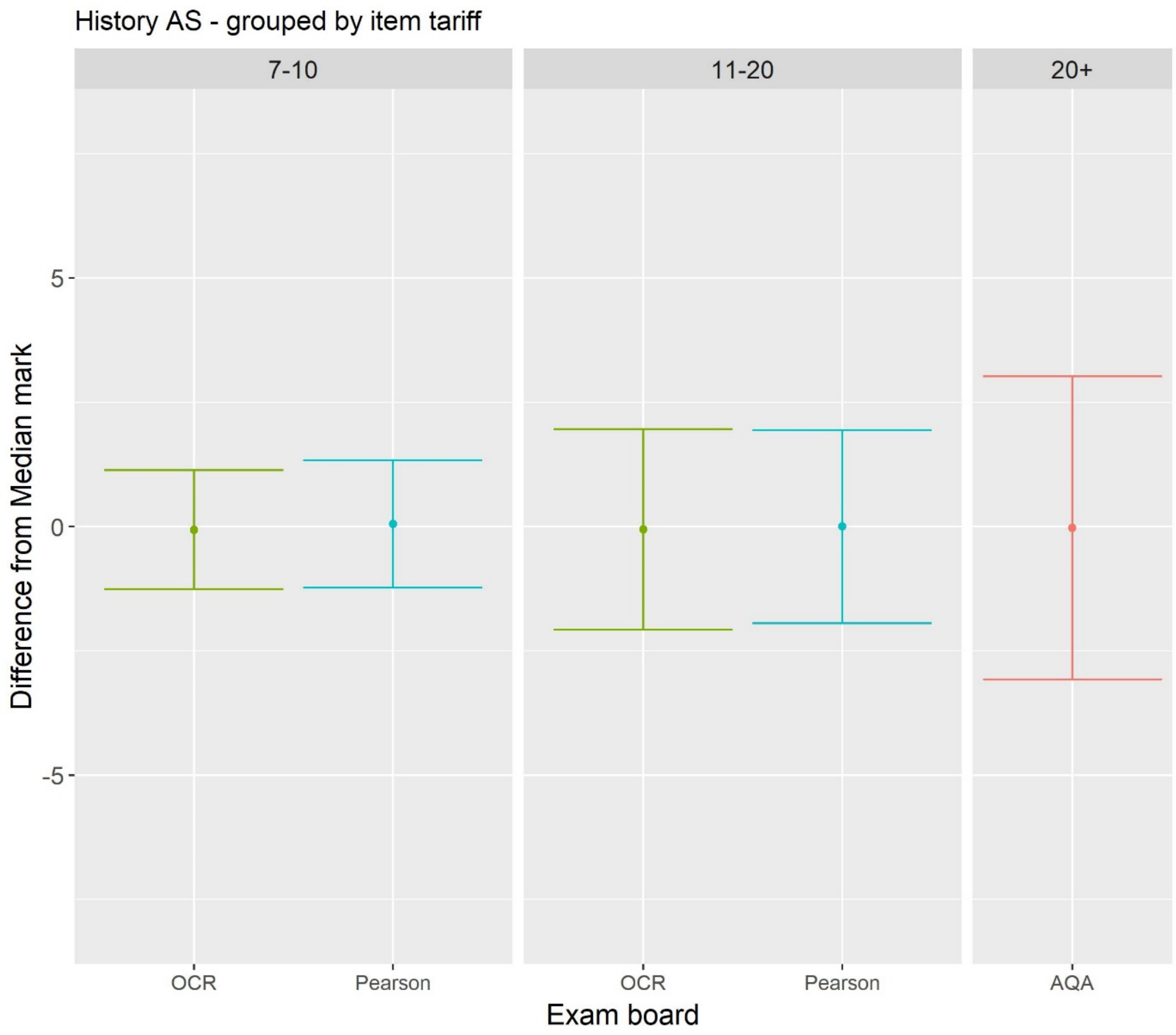


Figure D4: *Mean and standard deviation of the difference of item marks from the median mark, averaged across all markers and grouped by item tariff, for the summer 2016 history AS units. The exam boards are plotted separately and in different colours.*

2017 units

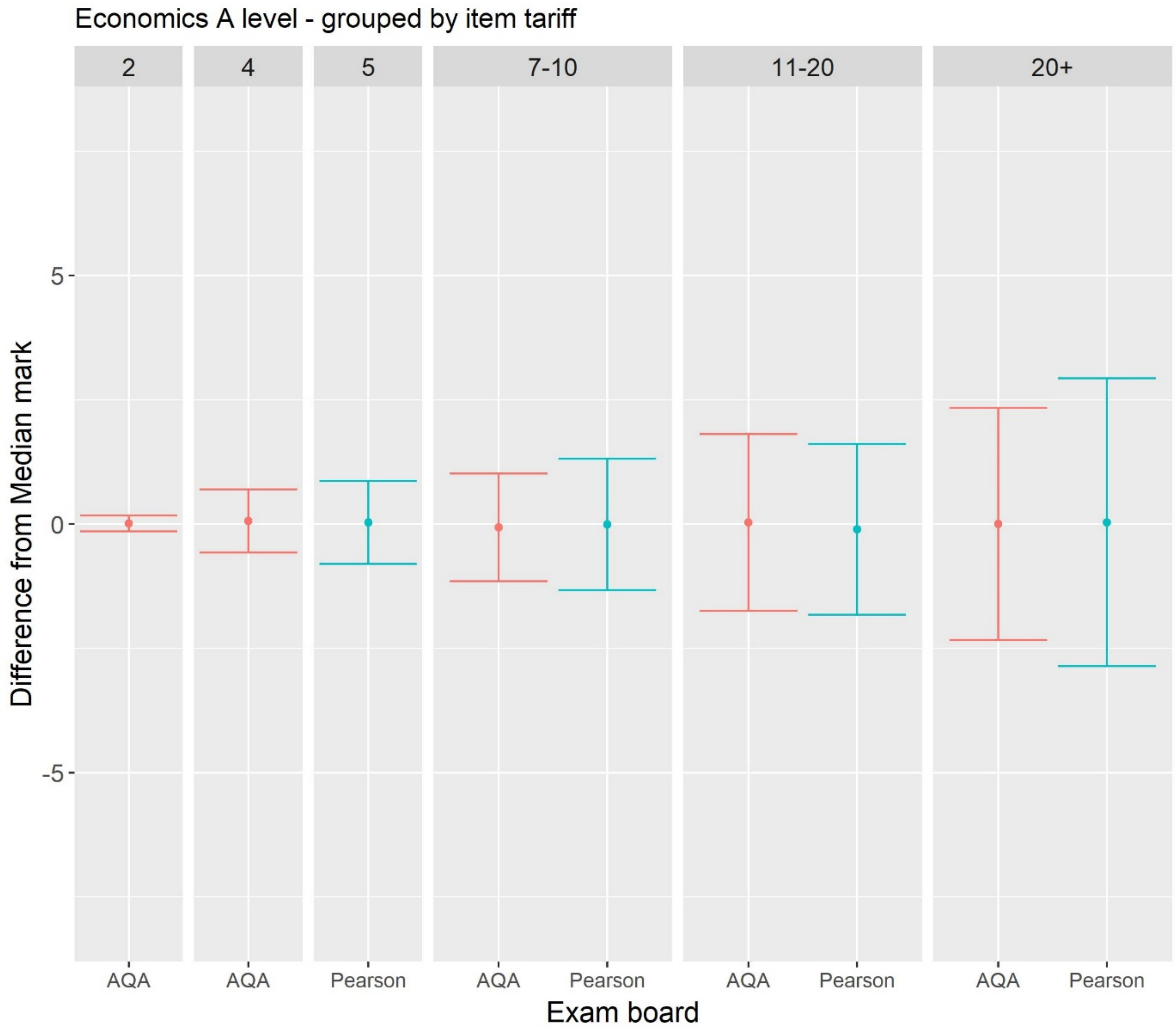


Figure D5: *Mean and standard deviation of the difference of item marks from the median mark, averaged across all markers and grouped by item tariff, for the summer 2017 economics A level units. The exam boards are plotted separately and in different colours*.

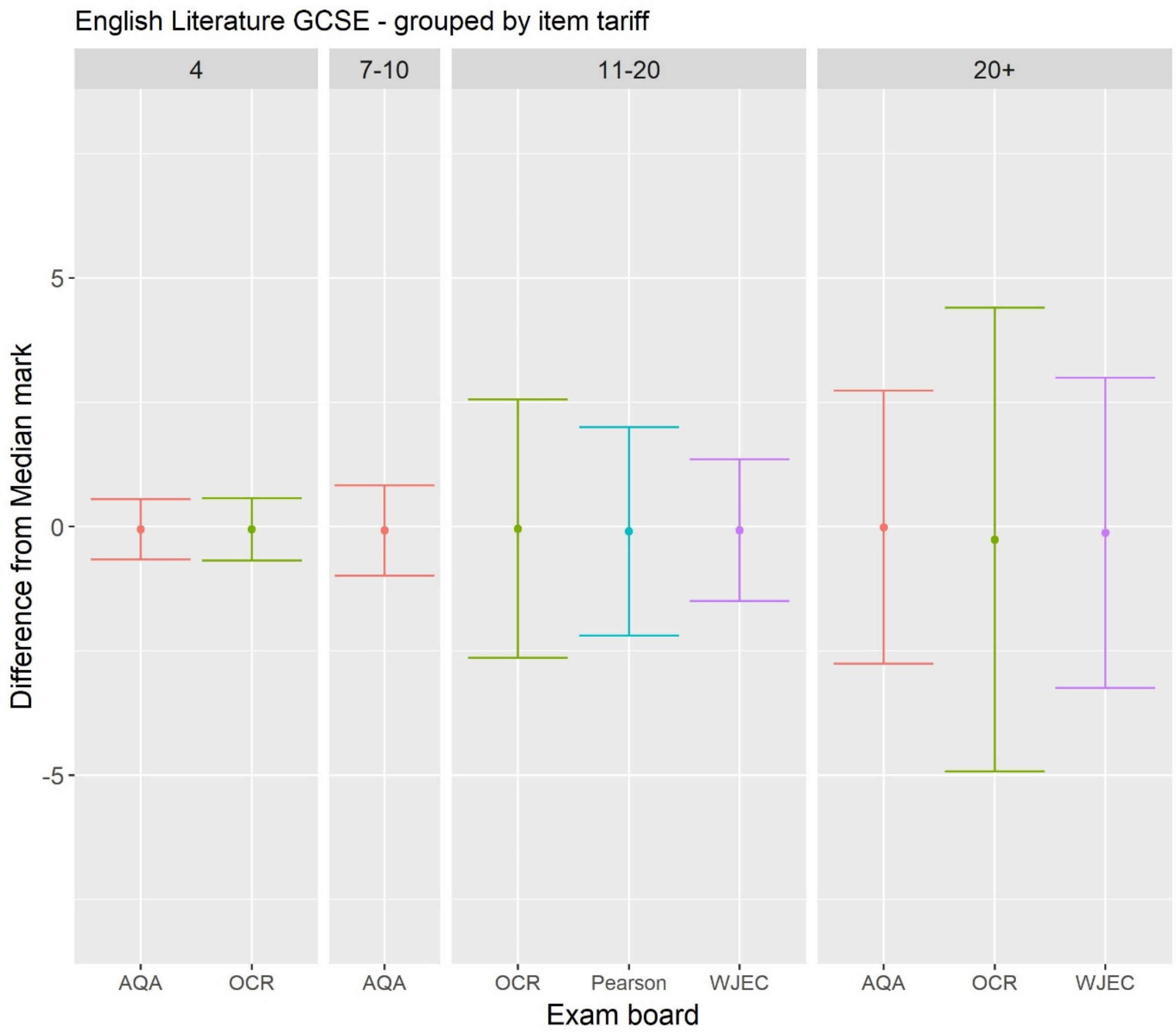


Figure D6: *Mean and standard deviation of the difference of item marks from the median mark, averaged across all markers and grouped by item tariff, for the summer 2017 English literature GCSE units. The exam boards are plotted separately and in different colours.*

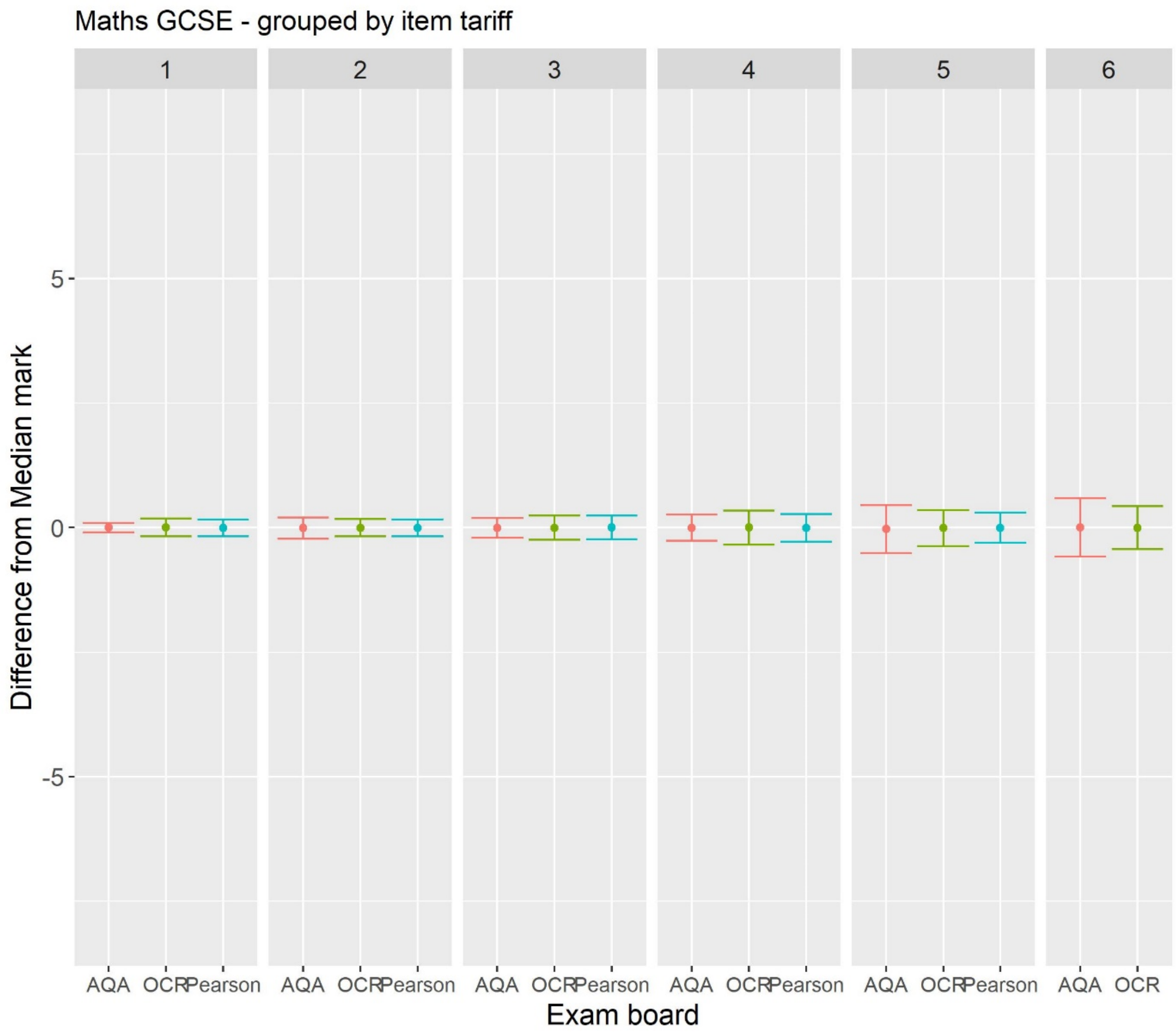


Figure D7: *Mean and standard deviation of the difference of item marks from the median mark, averaged across all markers and grouped by item tariff, for the summer 2017 mathematics GCSE units. The exam boards are plotted separately and in different colours.*

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

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

November 2018

Ofqual/18/6449/3