Research and Analysis

Marking consistency metrics

An update

ofqual

Authors
This report was prepared by Stephen Rhead and Beth Black from the Strategy, Risk and Research Directorate, and Anne Pinot de Moira, consultant.

Contents

Executive summary

A series of metrics are presented in this report. We have also reported on qualification level metrics for the first time, made possible since the introduction of fully linear qualifications and the removal of internal assessment from a number of subjects. It has been observed that qualification level metrics are generally higher than the components from which it is comprised. This report also describes on some work exploring the extent to which optionality within question papers can impact on component and qualification metrics.

Consideration of the practical uses of such metrics, such as the derivation of thresholds to identify acceptable minimum levels of marking consistency (and unacceptable) are also discussed. The report discusses how the determination of any thresholds should take into account the subject and/or assessment types in order to ensure the levels promote appropriate comparisons. In deciding which the most appropriate thresholds are, we should take into account both the number/proportion of components flagged, as well as the public acceptability of any threshold.

This report also shares some data (based on the marking of 6 mark items) comparing the marking in England with marking elsewhere in the world. This provides some indication that the marking in England is of similar levels of consistency as elsewhere. This report, using the same data, also indicates that marking consistency over time (between 2013 and 2017) appears to be relatively stable – it has neither deteriorated nor improved.

There is a range of values of metrics reported here, one of which is the probability of receiving the 'definitive' grade at qualification level. The term 'definitive'[1] is based on 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. And it should be noted that comparison to a single 'definitive' mark represents a relatively stringent measure of marking consistency. If other legitimate marks were to be modelled in[2], effectively changing the metric to the 'probability of receiving a legitimate mark' the probabilities would be somewhat higher.

As might be expected there are some clearly identifiable subject patterns. The probability of receiving the 'definitive' qualification grade varies by qualification and subject, from 0.96 (a mathematics qualification) to 0.52 (an English language and literature qualification). The probability of receiving the definitive grade or adjacent grade is above 0.95 for all qualifications, with many at or very close to 1.0 (ie suggesting that 100% of candidates receive the definitive or adjacent grade in these qualifications).

This is not to say that there are not components or qualifications where the marking consistency cannot be improved. Through identifying appropriate thresholds of acceptability, exam boards should channel additional resource and support to those

[1] Eg 'definitive' mark or 'definitive' grade

[2] The exam board systems only captures a single legitimate (definitive) mark for any response, so it is not possible to properly model in other legitimate marks.

components or qualifications which most need improving. Exam boards should, additionally, be looking for opportunities to incrementally improve marking.

All future work with metrics needs to proceed with some caution. This is to manage the risk that any use of thresholds or benchmarks do not compromise the live on-line monitoring procedures and hence the actual quality of marking, which is the very thing we wish to improve.

1. Introduction

In 2016, Ofqual published a report on marking consistency metrics (Rhead, Black, and Pinot de Moira, 2016), representing the first phase of work in deriving such metrics, and which was undertaken as a result of a previous recommendation (Ofqual, 2014). This work outlined different approaches for the derivation of metrics, limitations and assumptions and highlighted the potential impact these metrics might have on the live monitoring process. It also illustrated how these metrics could be scaled to qualification level and potentially used for linear qualifications, once the reformed GCSEs and A levels were phased in.

This current report provides an update to the 2016 report; using operational data arising from the 2017 summer session, and presents for the first time qualification level metrics which can be derived for the reformed GCSEs (GCSE 9 to 1) and A levels. A brief overview of the monitoring procedures employed by the exam boards, limitations and assumptions needed for the metrics will be given as will the derivation of component and qualification metrics. There is also a discussion on how metrics might potentially be used to establish acceptable levels of marking consistency for different assessment types. Finally, we also explore the extent to which optional questions, which may or may not be marked similarly to one another, and for which there may be different amounts of available data, may impact upon the calculation of the metrics.

2. The data behind the metrics

The data used for the derivation of the metrics is all sourced from the marker monitoring activities conducted by exam boards during live marking in order to quality assure the marking. This section briefly describes the exam board processes which generate the data as well as the subjects for which we collect the data.

2.1 Marking monitoring processes in live marking

All 4 exam boards (AQA, OCR, Pearson and WJEC) who provided marking data for this project use on-screen marking and monitoring for some components/units (referred to as components from here). This generates electronic records of the monitoring of quality of marking which we subsequently collect.

On-screen marking is mainly monitored using one of two procedures. The first and most common approach is the introduction of pre-marked responses into an examiner's script allocation. These pre-marked responses are known as seed or validity items (hereafter referred to as seed items). Seed items are introduced at times and intervals unknown to the examiner (sampling rates of approximately 5% are typical). The examiner is unaware that it is a seed item and marks the item without sight of the pre-determined mark. A comparison of the two marks derived from this process allows an assessment of the examiner's marking against a pre-agreed standard. This process is illustrated in Figure 1.

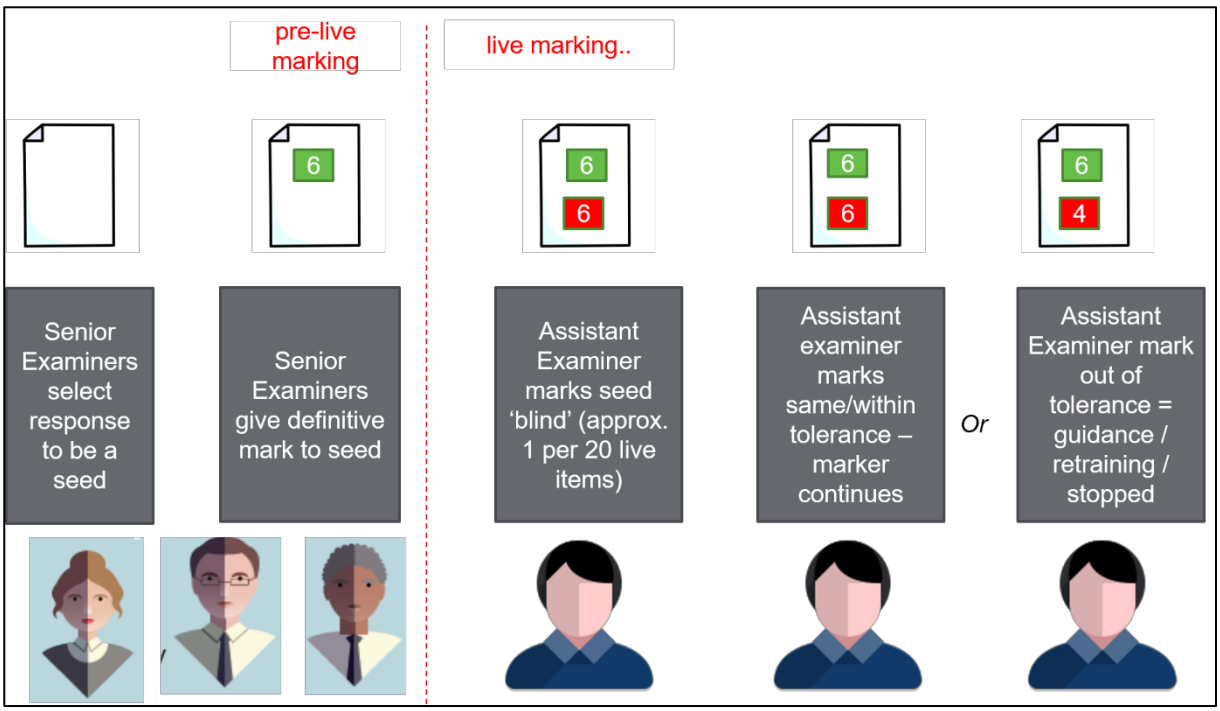


Figure 1. The seeding process. Prior to live marking senior examiners select responses to be seed items and assign a definitive mark to the seed item. The definitive mark awarded to a seed is that which will contribute to the final mark of a candidate. These pre-marked responses are introduced into an examiner's allocation at times and intervals unknown to the examiner. The mark awarded by the examiner does not contribute to the candidate's final mark and is used as a mechanism to monitor marking. If the examiner's mark agrees with the definitive mark or is within tolerance, the examiner can continue to mark. If the mark is out of tolerance the examiner may be given guidance or retraining, or stopped from marking.

All exam boards in this study have on-screen marking systems that allow monitoring by seeds. However, the exam boards have differing approaches to selection and distribution of seeds. Some boards and marking systems select and distribute seeds at item (question) level or small groups of items (questions), whereas some boards and marking systems select and distribute seeds only at the level of whole scripts. In this latter case, for any single examiner the seed is therefore the entire pre-marked script but item level information is still captured. In both systems (whole script or item seeding) the final mark for the seed item which contributes to the candidate's overall mark is known as the 'definitive' mark.

There are many ways for arriving at a single, definitive, mark of a seed item (see Tisi, Whitehouse, Maughan, and Burdett, 2013), although typically once the seeds have been selected, the definitive mark is generally derived by one or more senior examiners, often but not always including the Principal Examiner for that component. Exam boards generally allow some flexibility and there is no formal record for each seeding item of precisely who was involved in recording the definitive mark. In order to incorporate seed items in the derivation of marking consistency metrics it has been necessary to assume that the way in which the final mark is derived introduces no bias to potential of marking consistency metrics and to accept the seed mark as the definitive mark no matter how it was derived.

Along with seeds, some boards also employ a system of blind sample-double marking which is typically used for an extended response (illustrated in Figure 2). In this approach a series of randomly chosen responses will be blind marked by two randomly paired examiners. For all boards the examiners are chosen from the entire pool of examiners. However, how the final mark is awarded to the candidate varies by board. In one approach the final mark awarded to a blind sample double-marked response is the higher of the two marks unless they differ by more than a pre-agreed tolerance (the 'consensual approach'). For the second approach the second examiner is always a senior examiner and the final mark awarded is that of the senior examiner (the 'hierarchical approach').

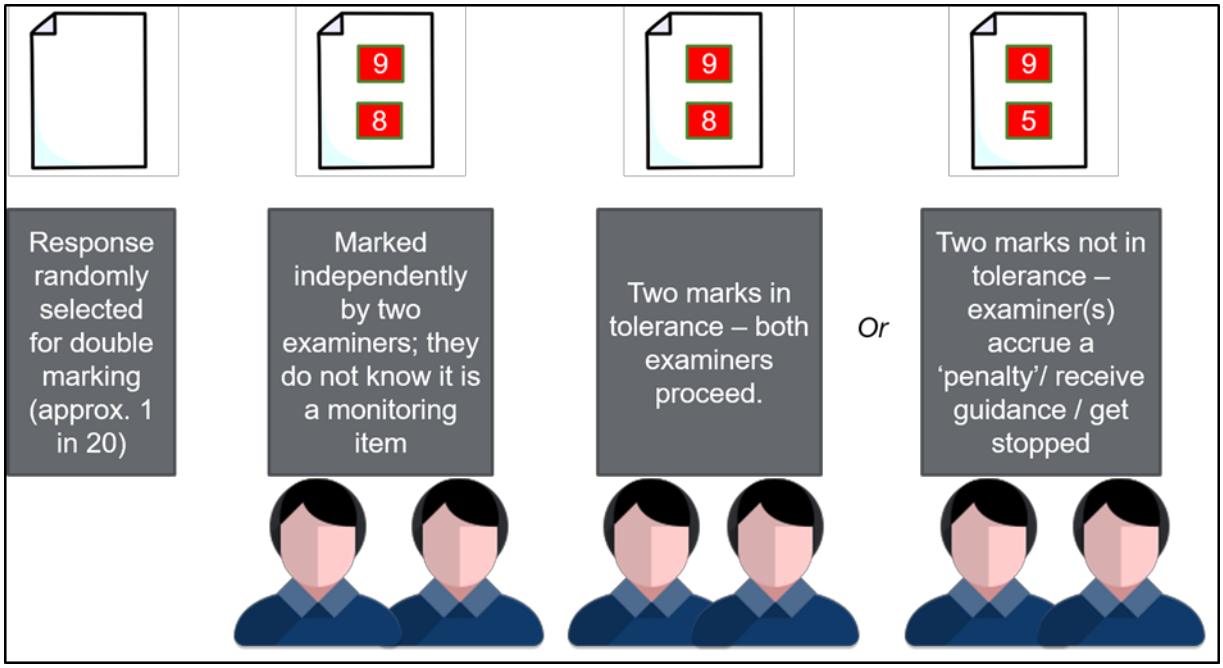


Figure 2. The process behind blind sample-double marking.

Regardless of either approach used (seed items or blind sample double-marking), the two marks awarded to a single response were arrived at independently of one other and as a result can be treated as independent in the statistical sense (Bramley and Dhawan, 2010).

2.2 Exam board data collected

The data collected is that generated from the operational monitoring of quality of marking during the live marking session. Data for all items on online marked components were requested from all 4 exam boards for the following subjects:

- Biology
- Business studies
- Chemistry
- Computer science
- Economics

- English language
- English language and literature
- English literature
- French
- German
- Geography
- History
- Mathematics
- Physical education
- Physics
- Psychology
- Religious studies
- Sociology
- Spanish

The data requested was for GCSE, AS and A level. The subjects represent the highest volume qualifications and also represent a range of item types and examination structures. Where both reformed and legacy qualifications were available, we collected only the data from the reformed qualifications.

This data set has 453 unique components and some 16.4 million 'marking events' of which approximately 16.2 million were generated from seed items and 198,000 were sample double marked items. Each item in the dataset has marks awarded by 2 or more examiners. This mark-remark data is the foundation of this analysis. For seed items the first examiner mark and the final mark awarded to the candidate are defined as the mark-remark data. Hierarchal sample-double marked items are analogous to this, the first examiner mark and the final mark awarded are defined as the mark-remark data.[3] The mark-remark difference is given by the following relationship:

$$difference = mark\ awarded\ by\ examiner\ 1 - final\ mark\ awarded\ .$$

A positive mark-remark difference means that the first examiner has awarded a mark more lenient than the definitive mark and negative difference corresponds to a more severe mark.

3. Metrics

In their 2010 report, Bramley and Dhawan present the idea of quality of marking as distinct from reliability of assessment, describing the concept as examiner-related variability or examiner accuracy. With this in mind, the metrics presented here are all derived from the mark-remark data arising from multiple responses to the seed and sample double-marked items.

[3] The final mark awarded to the item was missing for some consensual sample-double marked data. In such cases the first examiner mark and second examiner mark were defined as the mark-remark data.

Ideally marking consistency metrics should be presented at the least granular level possible allowing comparisons between qualifications in the same subject area. With the move from modular to linear assessment and a reduction in non-examined assessment as features of the reformed GCSEs and A levels, it is possible to derive qualification level metrics by the aggregation of marking consistency metrics for all components within a qualification. As the majority of on-screen marking is segmented (ie distributed at item level rather than script level) and a single candidate's work is not used as a seed across all components it is not possible to track how an examiner would mark all components for a single candidate within a qualification. As a result, qualification level metrics are derived from item level statistics for each question within a component (Figure 3). For a more in depth discussion on these metrics please refer to section 4.2 in Ofqual's 2016 report (Rhead et al., 2016).

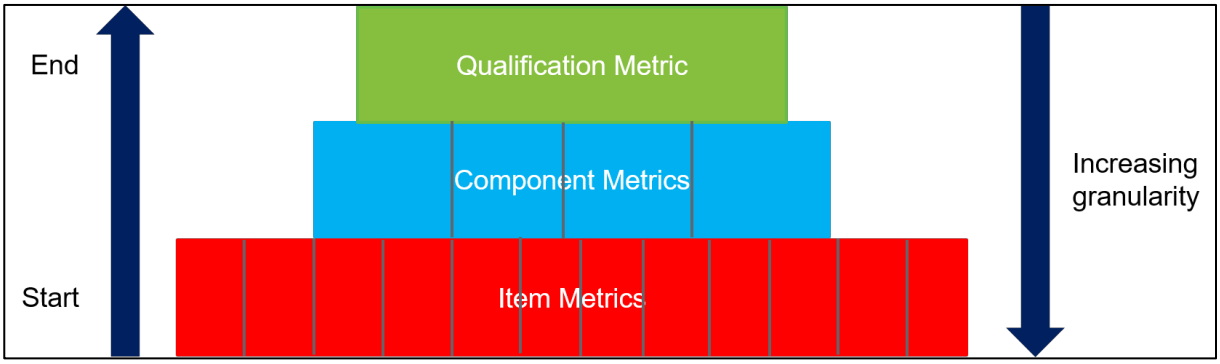


Figure 3. The process of deriving component and qualification level consistency metrics within a single qualification. Component level metrics are derived by the aggregation of item level statistics for all questions within a particular component. Likewise, qualification metrics are derived by aggregating overall components in a qualification. These statistics are complementary; a metric from one level may be used to contextualise information in another.

3.1 Limitations and assumptions

A series of component and qualification level metrics are derived from exam board data arising from on-line monitoring procedures. It has been necessary to make a series of assumptions in the derivation of these metrics. All the analysis in this report has assumed that the most appropriate basic measure of consistency of marking is the difference between 2 independently awarded marks. In order to use the data from the exam boards, it has been necessary to assume:

1. That the mark awarded to the seed item is the definitive mark. This is most likely the case for most seed items, but in instances where the most frequent mark awarded by examiners differs from the definitive mark there is a possibility that the definitive mark is wrong (Bramley and Dhawan, 2010). There are multiple approaches used for arriving at the definitive mark (Tisi, 2013) and, as there is no formal procedure for arriving at a single mark for a seed item, nor is there any formal recording of the process, it has been

necessary to assume that no bias is introduced to the potential of marking consistency metric by the way in which the final mark is derived.

2. That the 2 marks being compared are entirely independent. The assumption of independence is safer perhaps for sample-double marking than for seeding items. In seeding, for example, it is possible in some marking systems that in some cases those examiners involved in deriving the definitive mark for seeds are subsequently monitored using the same seeds and that they may remember some of the marks awarded. Additionally, it may be that examiners receive feedback (including the mark) on specific seeds and are able to retain and re-use this information if the same seed reappears subsequently. Where 2 marks are not independent, this would most likely provide an over-estimate of marking consistency for the purpose of these metrics[4].

There are other assumptions present in how metrics have been derived, including how to deal with optionality[5] within question papers. It has been assumed that it is acceptable to 'collapse' optional questions in the derivation of component and qualification level metrics.[6] This appears to be a reasonable assumption when comparing the distribution of differences between simulated and actual values in the 2016 report. However, this could be problematic if marking consistency between optional questions is very different. It is also observed that there are not seeds for all optional questions and, using the limited data available on weightings of optional questions within a component, therefore it is possible that the seeds are not entirely representative of the numbers of candidates taking each route. Ideally, each optional route through a component or qualification should be treated as a separate entity (Stockford and He, 2014). However, this will be difficult to do due to missing questions and weightings. This is discussed in section 3.6 and the appendix.

Because the metrics are derived directly from seeding data from monitoring, the metrics can only reflect those seeds and their data; one potential limitation is worth pointing out. For example, if an item is worth 10 marks and the chosen seeds represent a range in marks from 3 to 7, then this is reflected in the mean difference for that particular item, which in turn is reflected in the metrics. Ideally, if on-screen marked data is to be used in the derivation of metrics, seeds should be selected across the entire mark range of the item, including zero and full-mark responses.

It would be very easy to artificially improve the metrics by only including 'easy' seeds into the monitoring process – ie only testing very straightforward to mark responses, rather than including a mixture of responses (likely to be representative) that include, as well as straightforward to mark, the less straightforward to mark. There is an assumption in the metrics that, broadly speaking, the seeds have provided a 'fair

[4] It is also worth pointing out that any loss of independence of the two marks not only undermines the metrics but also undermines the true purpose of seeding which is to monitor live marking.
[5] Optionality means that some or all of the questions are optional. For example, candidates might have to answer all questions in section A, and choose one question from say 6, from section B. Subjects where optionality is most common are English language, English language and literature, English literature, history, religious studies and sociology.
[6] ie combine all seed data from within an optional set of questions, as if they were derived from a single question.

test' of marking for the markers – not too easy, not too difficult, representing the scope of the nature of responses and the challenge normally encountered in marking. If seeds were chosen for responses that were particularly difficult to mark, (for example, to check the examiners' understanding of how to apply the mark scheme in atypical responses) and these seeds were over-represented, then these metrics would under-estimate the consistency of marking for non-seed items (Bramley and Dhawan, 2010).

Due to the majority of on-screen marking being segmented at item level, there are some questions in components that have no mark-remark data. This happens for one of two scenarios: the questions were automarked and not included in the mark-remark data, or the questions are optional and have no seed items. Automarked items are typically multiple choice or objective response items which can be computer-read. In order to derive component and qualification metrics it is necessary to reintroduce the missing automarked questions into the dataset. This is done by assuming that the missing automarked questions are marked perfectly accurately. Missing optional questions are assumed to have the same marking consistency as the other optional questions when collapsed into a single question. Analysis in this report focusses on components where responses to all questions were present.

3.2 Sum of independent random variables

For each question within a component it is possible to calculate the mean and standard deviation of the difference from the awarded mark. From this an estimate of the mean and standard deviation at component level can be obtained. If $\mu(X_1)$, $\mu(X_2)$, … , $\mu(X_n)$ are the mean difference for each of the n questions within a component and $X_1, X_2, \ldots, X_n$ are random variables with known distributions, the mean difference from the awarded mark at component level may be given by:

$$\mu(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} \mu(X_i). \quad (1)$$

Likewise, the variance at component level can be expressed by:

$$V(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} V(X_i) + \sum_{i \neq j} \operatorname{cov}\left(X_i, X_j\right) \approx \sum_{i=1}^{n} V(X_i), \quad (2)$$

where $V(X_n)$ is the variance of the n^{th} question within a component. Due to the segmented nature of the majority of on-screen marking, it has been assumed that the distribution of differences between questions are independent, as a result the covariance term is zero. The standard deviation is obtained by taking the square root of equation 2.

The mean difference and standard deviation at qualification level are estimated from all items within a qualification in an identical approach to that at component level (equations 1 and 2). Any qualifications with missing data from one or more external components are excluded from analysis. Similarly, any qualifications which include internal assessments are also excluded as there is not the same monitoring data available for these components. Details on the number of components and qualifications for each subject are presented in Table 1.

Table 1. Breakdown of components and qualifications for each subject

Subject	Number of components with monitoring data	Number of components with monitoring data which contribute to the full qualification level analyses	Number of qualifications with full component data
Biology	39	14	8
Business studies	15	13	6
Chemistry	37	13	7
Computer science	6	0	0
Economics	19	14	7
English language	19	14	7
English language and literature	11	6	3
English literature	27	18	10
French	14	0	0
Geography	31	6	3
German	12	0	0
History	64	9	6
Mathematics	60	18	6
Physical education	5	0	0
Physics	45	24	11
Psychology	15	15	6
Religious studies	13	8	3
Sociology	7	7	3
Spanish	14	0	0
Total	453	179	86

Equations 1 and 2 allow an estimate of consistency of marking in terms of the mean difference and standard deviation from the definitive mark at component level. So we can compare components with different maximum marks, items etc, we scale the mean difference and standard deviation by the maximum mark of the component. This can be seen in Figure 4 for all physics components, where typically (with just one exception) the mean difference and standard deviation are within ± 2% and ± 4% of the maximum mark of the component respectively. As the mean difference is close to 0%, suggesting that examiners for each component show no systematic bias towards severe or lenient marking. Marking consistency is generally similar for most physics components regardless of board or level.

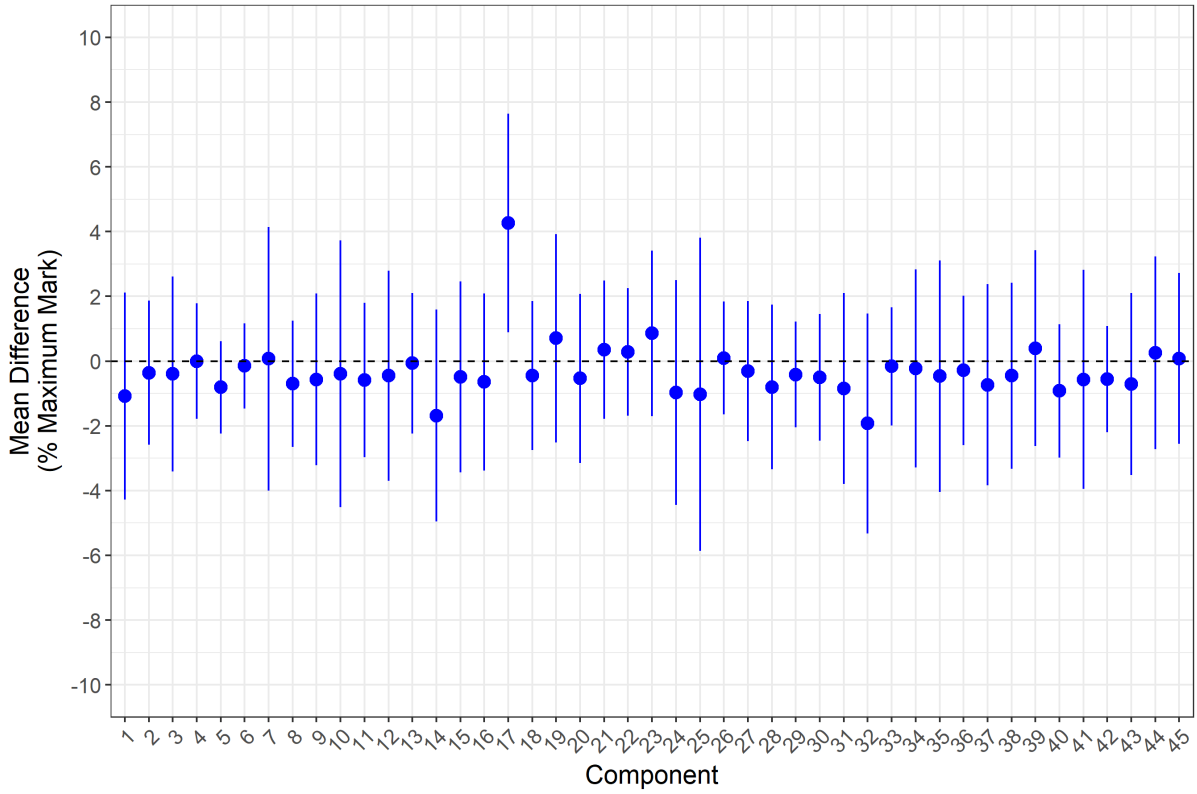


Figure 4. The mean difference (solid circle) and standard deviation (whiskers) of the 'definitive' mark expressed as a percentage of the maximum mark for each physics component.

Figure 5 shows the mean difference and standard deviation from the definitive mark at qualification level for the 11 AS and A level physics qualifications with no externally assessed components (no GCSE physics qualifications are reported on as physics GCSE was not reformed in 2017). Marking consistency is reasonably similar between all qualifications; the mean difference and standard deviation are typically found to be within ± 1% and ± 2% of the maximum mark of the qualification respectively.

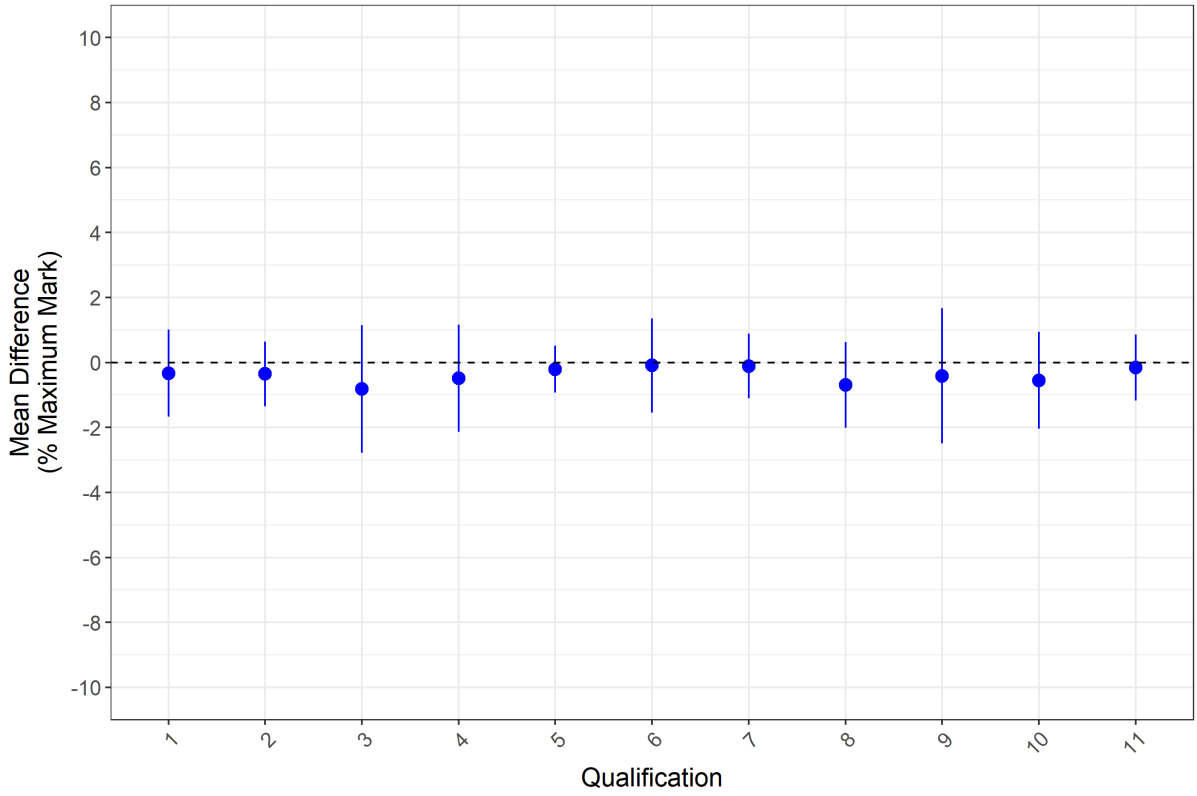


Figure 5. The mean difference and standard deviation of the ‘definitive’ mark expressed as a percentage of the maximum mark for each physics qualification.

3.3 Probability of ‘definitive’ grade

As identified in the 2016 report, components or qualifications need to be on a common scale to allow for meaningful comparisons; the grade scale is one such common scale. We have developed a metric which relates consistency of marking to the grading. The approach is described in Rhead et al (2016). In short, the unscaled mean difference and standard deviations are used to calculate the probability of a particular mark resulting in the ‘definitive grade’, ie the same grade classification that would result from the definitive mark for the seed item. The probability of a particular mark resulting in the definitive grade classification can be calculated using the distance to the nearest grade boundaries as cut points on the normal distribution (Figure 6). For each mark the black line represents the probability that the candidate has been awarded the definitive grade. The probability dips in the mark region near the grade boundaries and is highest at the extremes of the mark distribution. To a large extent, the probability that a candidate is awarded the definitive grade is determined by the mark position relative to the grade boundary; a script where the mark is on or near the grade boundary but which is marked severely or leniently by a single mark is at a greater risk of not receiving the definitive grade compared to a script several marks away from a grade boundary or comfortably within grade. For any qualification, the probability of receiving the definitive grade is significantly influenced by the overall spread of the grade boundaries. In components or qualifications where the grade boundaries are close together (most likely because the assessment has not successfully spread out candidate marks) the marking consistency will have more of an impact on the probability of being awarded the

definitive grade. Thus, the wider the gradewidth, ie the wider spread of grade boundary locations, the greater the probability of being awarded the definitive grade. Different components and qualifications have different gradewidths. This is largely a function of how successfully the assessments have spread out the marks of candidates of differing abilities, ie a product of good assessment design, rather than the number of grades available for the qualification. In other words, given that item writers and test designers know the number of grades for the qualification in advance, assessments should be designed such that they will differentiate reliably across the gradeset.

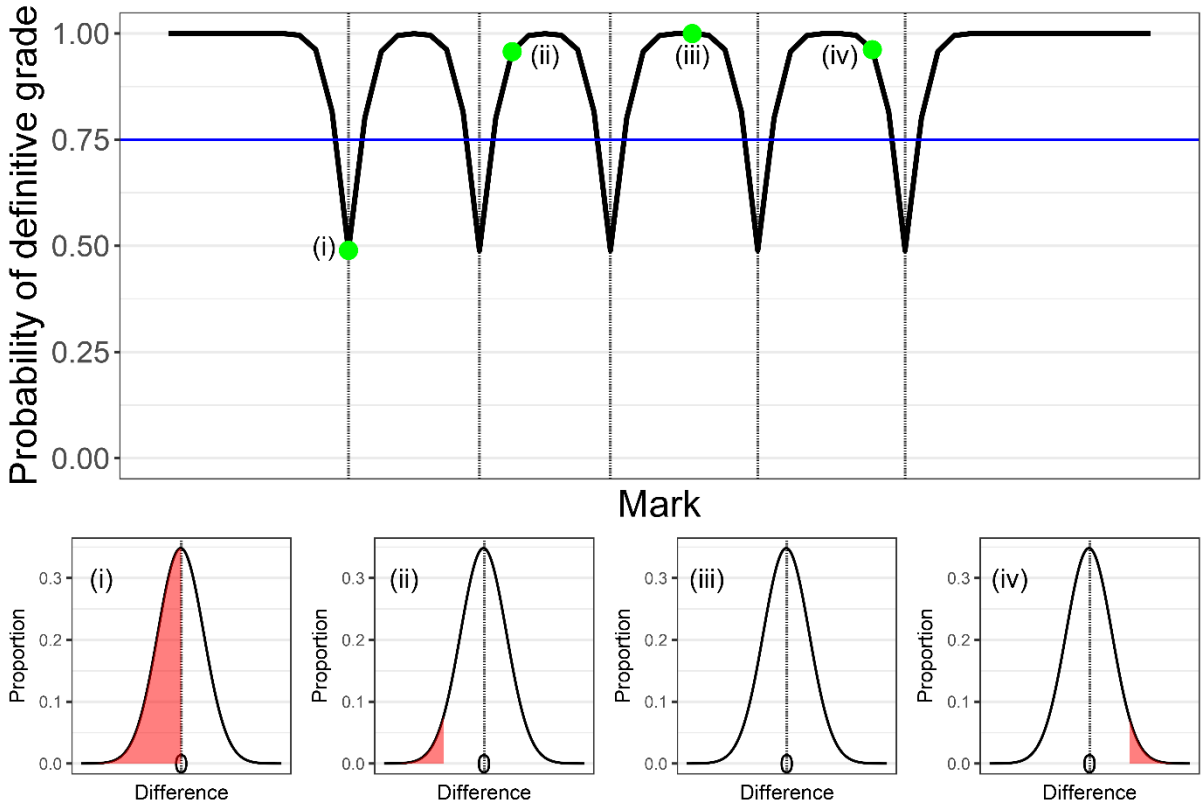


Figure 6. The probability of being awarded the 'definitive' component or qualification grade dependent on the final mark awarded to the candidate (solid black line) for an AS component or qualification, where grade boundaries are given by the dashed line. The probability at each mark is calculated from the proportion of candidates that are over- or under-graded and is illustrated by the shaded regions at various points ((i), (ii), (iii) and (iv)). The mean probability weighted by the mark distribution is given by the solid blue line.

A summary statistic of the overall probability of receiving the definitive grade is calculated by taking the weighted mean of the probability that a candidate has been awarded the definitive grade. In order to make these probabilities as reflective of the actual cohort that has taken the qualification, the mean is weighted by the mark distribution for the particular component or qualification (NB: this is different from the 2016 metrics report, where mark distributions were simulated). The range of summary statistics, for all components for which we have mark-remark data, presented by subject, is presented in Figure 7 below.

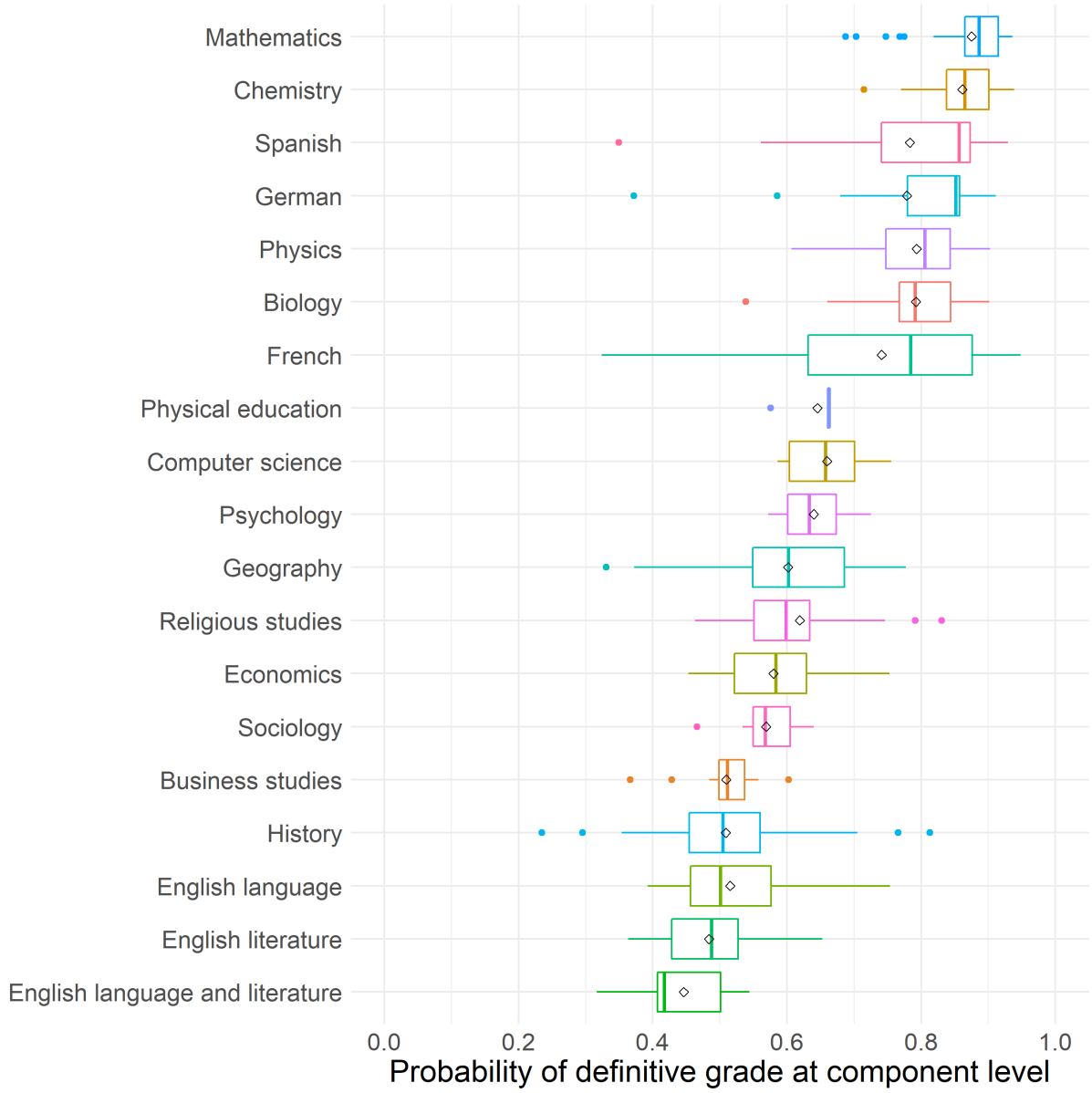


Figure 7: Boxplot of the probability of a candidate being awarded the 'definitive' grade at component level, by subject. The mean probability is denoted by the black diamonds. The total number of components is 453 overall and for each subject can be seen in Table 1.

3.4 Aggregation to qualification

As explained in previous sections, it is now possible to develop qualification level metrics by the aggregation of marking consistency metrics for all components within a qualification. The mean differences and standard deviations at qualification level are estimated from all items within a qualification (equations 1 and 2). Such an approach is illustrated below for a reformed higher tier GCSE mathematics qualification and an AS history qualification.

In general, we can expect that the probability of receiving the definitive grade at qualification level is greater than the probability of receiving the definitive grade for its constituent components. A worked example is included to help illustrate this.

Figure 8 shows a comparison of the scaled mean difference and standard deviation from the definitive mark at component and qualification level for a reformed foundation tier GCSE mathematics qualification. The mathematics qualification is comprised of 3 components all of which have the same maximum mark. Across the 3 components, examiners are typically found to award a mark slightly more severely or leniently than the definitive mark (the mean difference is within ± 0.5 %). In these components, marking is very consistent (the standard deviations are all within ± 2%). It is observed upon aggregation from component to qualification level that component level differences are averaged out. The scaled standard deviation at qualification level is smaller than those of the individual components. This means the qualification marking, overall, is more consistent at qualification level than at component level. Statistically, this occurs because although the maximum marks of the components combine linearly, the standard deviation does not (recall from equation 2 that the standard deviation is calculated by taking the square root of the summed variances for each item within a qualification), which leads to a reduced standard deviation as a percentage of maximum mark. From a common sense perspective, it makes sense that when combining the different component metrics to reach a qualification metric that the mean differences generally average out and, overall, the variability decreases because the mark differences (positive/negative) tend to cancel out when aggregated.

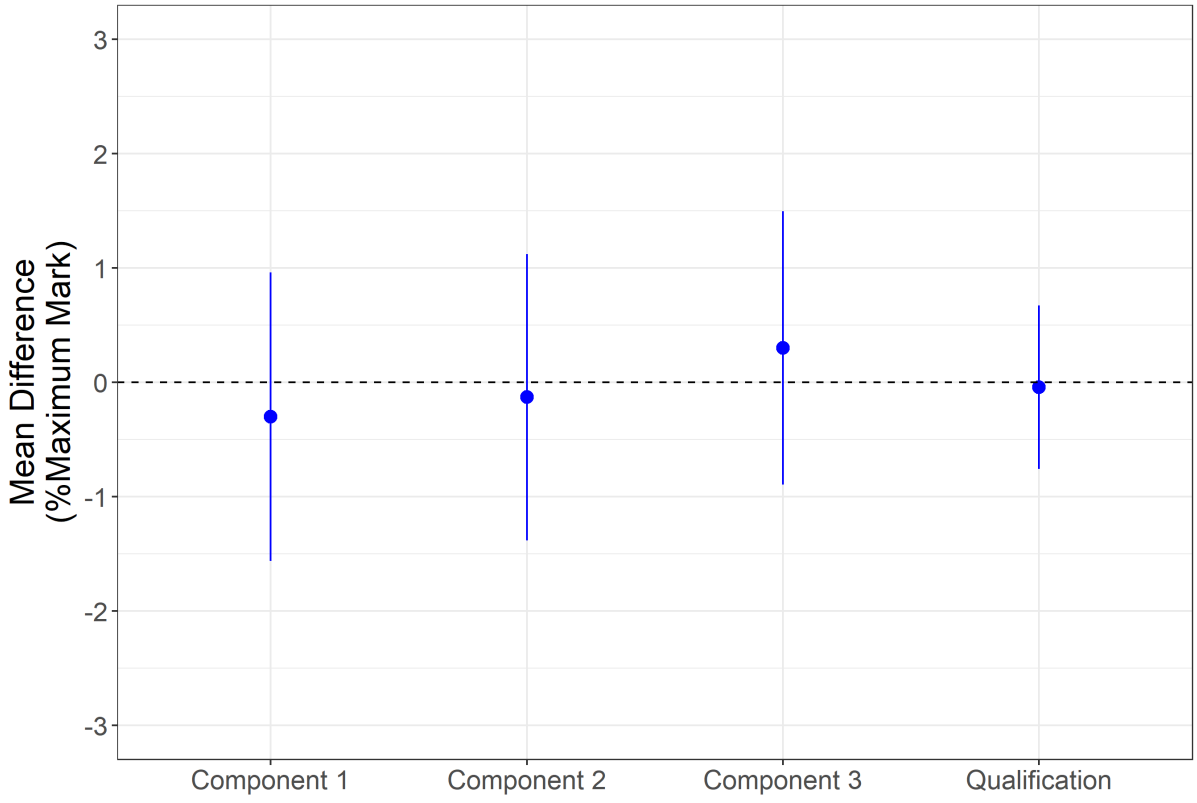


Figure 8. Comparison of the scaled mean difference and standard deviation from the 'definitive' mark at component and qualification level for the foundation tier reformed GCSE mathematics qualification.

Figure 9 shows a comparison of the probability of being awarded the definitive grade at component and qualification level for the higher tier reformed GCSE mathematics

qualification. Another feature of linear qualifications is that the component grade boundaries combine linearly (notwithstanding component weighting) to create the qualification grade boundary. The increase in the grade boundary widths, coupled with the decrease in the standard deviation as a function of maximum mark, means that the probability of achieving the definitive grade at qualification level is higher than that of the constituent components. It is observed that, apart from a small number of marks either side of the grade boundaries, the probability of a set of marked scripts getting the definitive grade at qualification level is 100%.

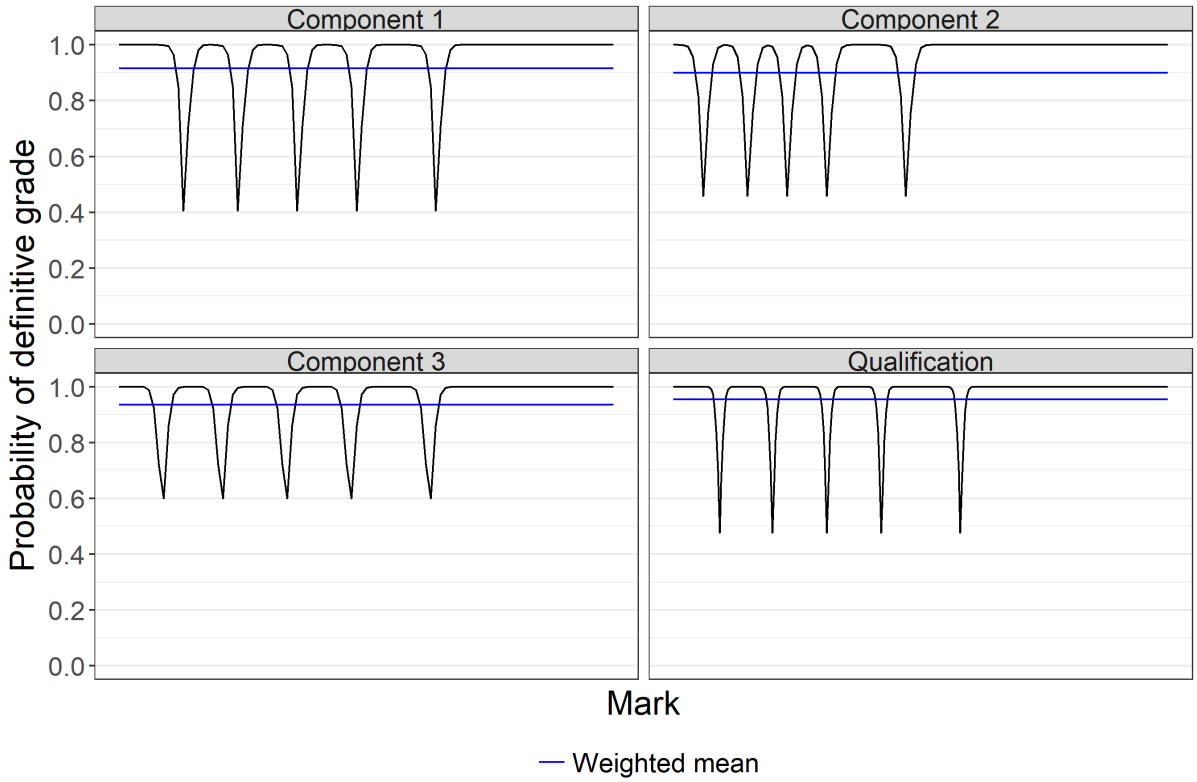


Figure 9. Comparison of the probability of achieving the 'definitive' grade at component and qualification level for the foundation tier reformed GCSE mathematics qualification.

Similar observations are made for an AS history qualification (Figure 10 and Figure 11). Examiners typically award a mark more lenient than the definitive mark and marking is more consistent at qualification level (as seen by the reduced standard deviation as a percentage of maximum mark). The probability of achieving the definitive grade at qualification level is again higher than those of the constituent components.

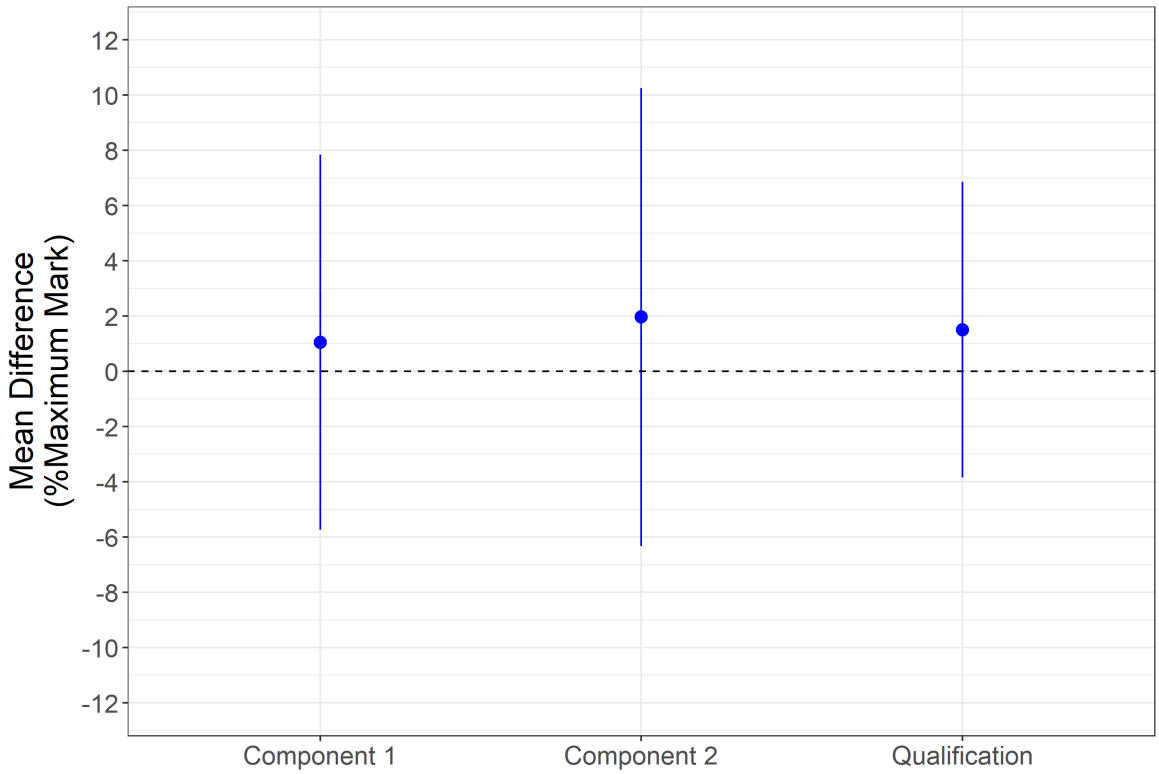


Figure 10. Comparison of the mean difference and standard deviation from the 'definitive' mark at component and qualification level for the AS history qualification.

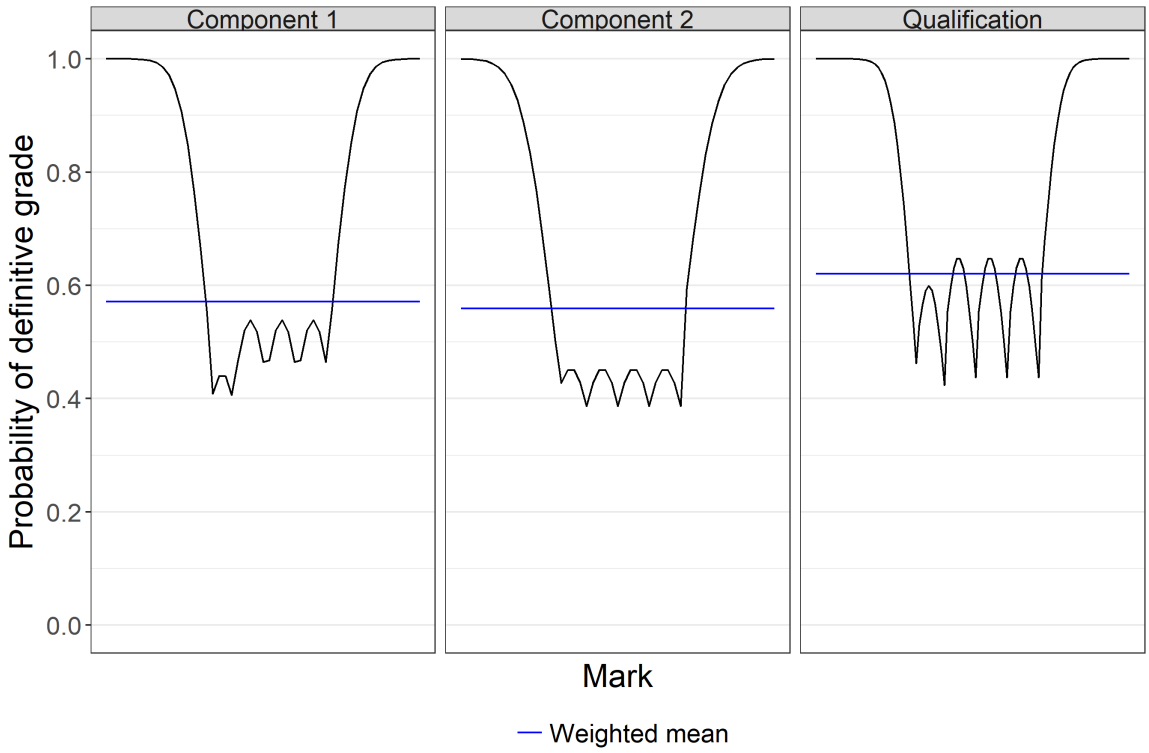


Figure 11. Comparison of the probability of achieving the 'definitive' grade at component and qualification level for the AS history qualification.

A comparison of the probability of being awarded the definitive grade at component level between subjects is given in Figure 12. There are different levels of marking

consistency in different subject areas. For example, consistency of marking for mathematics components and qualifications is higher than that for more 'subjective' English language components. Mathematics questions are generally low mark tariff questions with an objectively correct answer, whereas for more subjective questions, there may be some legitimate differences in the marks awarded in applying mark scheme between different examiners resulting in less agreement between examiners. As noted earlier the comparison to a single 'definitive' mark represents a relatively stringent measure of marking consistency and that if other legitimate marks were able to be modelled in[7], the value of these metrics would rise. Any future comparison between marking metrics should therefore only be between closely related subjects, or assessments with very similar item types.

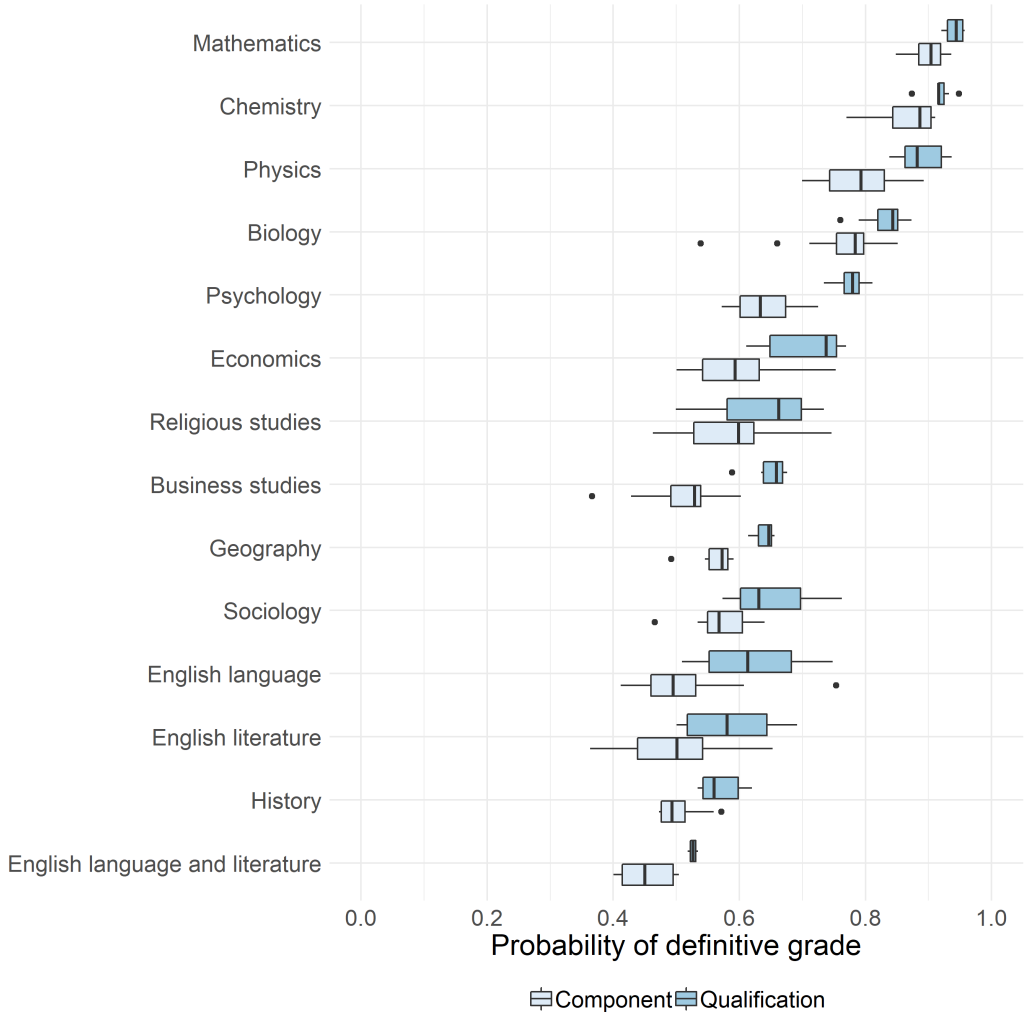


Figure 12. Boxplot showing the comparison of the probability of being awarded the 'definitive' grade at component and qualification level, for those GCSE, AS and A level qualifications for which we have full component data.[8]

[7] It is not possible to model any other legitimate marks in as the data contains only the definitive mark; where there might be other legitimate marks, the value of these are not captured in the exam board systems and are therefore unknown.

[8] In this graph, we can only aggregate up to qualification level metrics were we have data from all components. Accordingly, only those components which are part of qualifications for which have all data are included in this graph. This means (a) that some subjects are not present in this graph, and

It is also observed from Figure 12 that the probability of being awarded the definitive grade always increases upon aggregation from component to qualification level. This is due to an increase in grade boundary width and reduction of the scaled standard deviation.

Analysis can be extended to calculate the probability of a candidate being awarded a grade within one grade (ie +/- 1 grade) of the definitive grade. This is illustrated for all subjects in Figure 13. It is observed that the probability of being awarded a grade within one grade of the definitive grade is 1 or nearly 1 (ie 100% or near 100% probability) for nearly all subjects. For English literature, history, English language and literature, geography, sociology and English language, the median probability ranges from 0.96 to 0.99. However, it is possible that this might give an overly optimistic view of marking consistency because the extreme grades, ie the highest and lowest grades (eg A* and U respectively for A level), are associated with high levels of marking consistency because they are extremes, and small mark differences make no impact on the probability of being awarded the definitive grade. This can be seen clearly in Figure 11 where the extreme ends of the plot of the probability according to mark are very high. Pilliner (1969) (cited in Cresswell, 1986) formulated the aspiration that 95% of all candidates with a particular grade (other than the highest or lowest grade) should have the 'true scores' either in the grade they are given, the one above or the one below. He, along with others (Please, 1971, Mitchelmore, 1981) reason that unless an examination is perfectly reliable[9], some of those who receive marks just within one grade, will have 'true scores' that fall the other side of it. This is interesting, because although these arguments are addressed at wider concepts of reliability, the principle is still the same. The identification of an arbitrary benchmark of 95% is helpful for this more conservative estimate. Figure 14 shows the probability of being within ± one grade of the definitive grade, for all grades except top and bottom grades[10], for GCSE, AS and A level. This shows probabilities very similar to Figure 13, ranging from 0.95 to 0.99. One reason why the levels of probability of definitive grades are so similar between the 2 graphs, is that by excluding the extreme categories, relatively few students have been excluded, in other words, the overall probability estimate, in being based on the underlying mark distribution, is relatively stable when removing the 2 extreme grades.

(b) not all components for which we have data are included in this figure. All component data is presented earlier, in Figure 7.

[9] In the wider sense of reliability, incorporating marking reliability.

[10] For all levels, this means excluding the probability of achieving Ungraded (U). It also means exluding the highest grade as follows: for GCSE we excluded grade A*; GCSE 9 to 1 we excluded grade 9; for AS we exclude A; for A level we excluded A*.

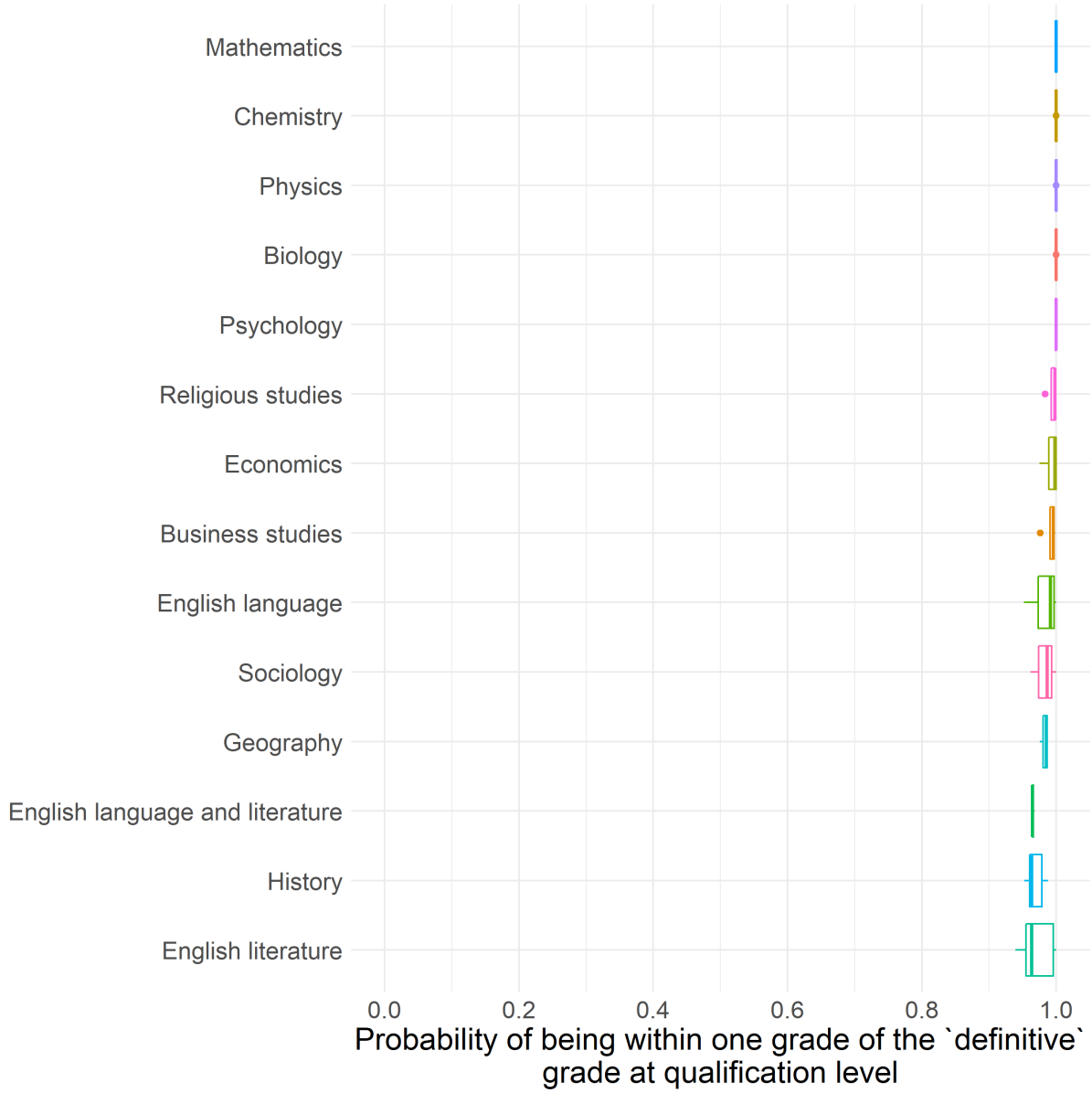


Figure 13. Boxplot, by subject, showing the probability of being within ± one grade of the 'definitive' grade, for all grades, for GCSE, AS and A level.

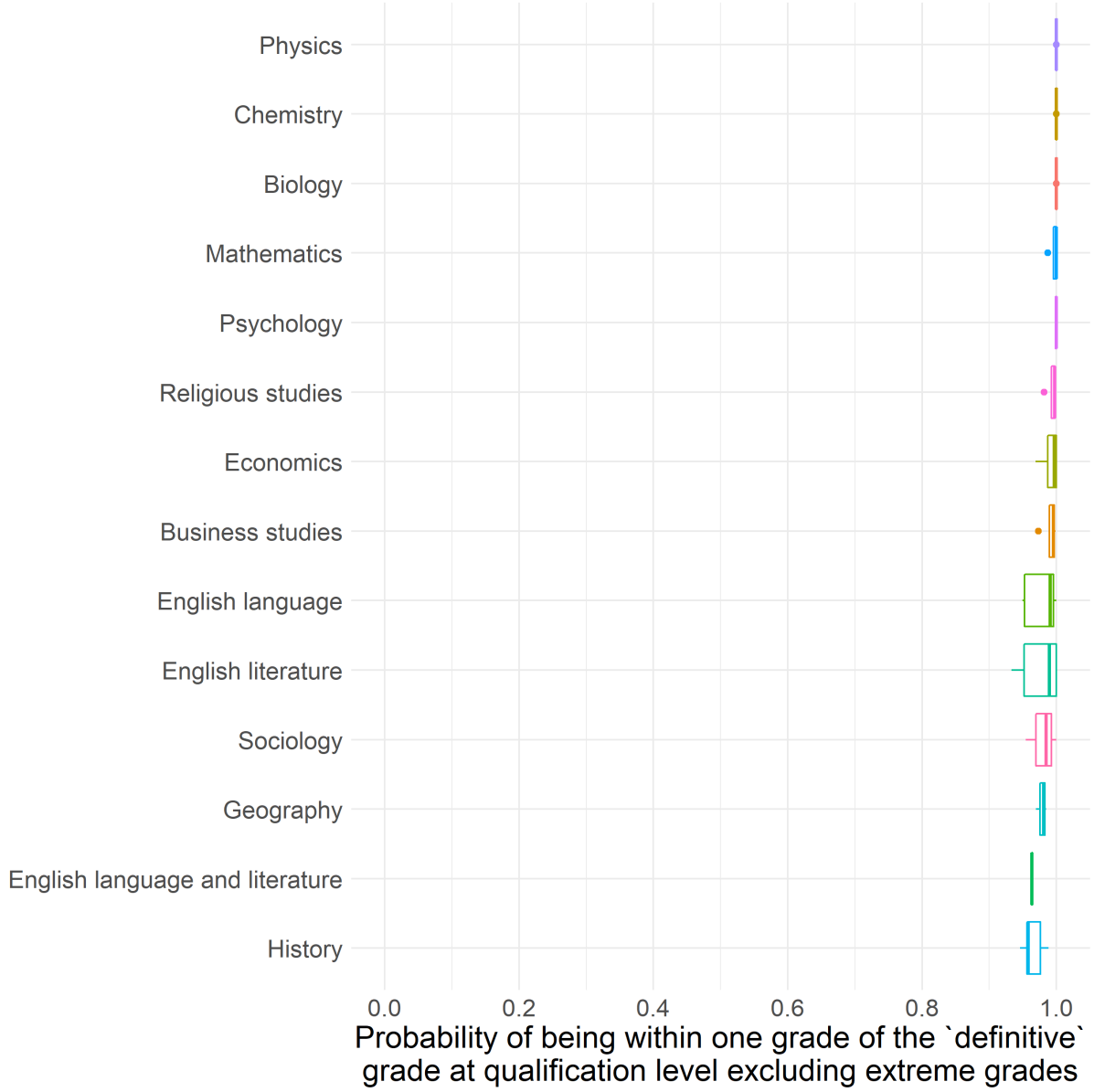


Figure 14. Boxplot, by subject, showing the probability of being within ± one grade of the 'definitive' grade, for all grades except top and bottom grades[11], for GCSE, AS and A level.

3.5 Marking consistency compared to international benchmarks and over time

In order to understand how well marking consistency in England compares to elsewhere, we undertook a short review of the international literature. The best point of comparison is the most granular, ie item level, on the basis that all other metrics build upon these.

[11] For all levels, this means excluding the probability of achieving Ungraded (U). It also means exluding the highest grade as follows: for GCSE we excluded grade A*; GCSE 9 to 1 we excluded grade 9; for AS we exclude A; for A level we excluded A*.

It is worth bearing in mind that the items, mark schemes and marking processes included in the international marking studies might not be exactly like those in England, so we are not strictly comparing like with like. However, it still provides some sort of helpful comparison or contextualisation for England. Many studies in the literature included 6 mark items, and the exact agreement rates varied from 46% (.46) to 75% (.75). Longford (1996) looking at essays included in a test of written English found exact agreement with a definitive mark ranged from 65-75%; Penny, Johnson and Gordon (2000) found that essays scored on 6 levels using a two phased approach gave exact agreement rates in the range of 59-63%. On items with slightly higher tariffs, similar rates of agreement were seen. For example, Brown, Glasswell and Harland (2004) found agreement rates between 66% and 92% for essays scored from 1-11; and Supovitz, MacGowan and Slattery (1997) found that reading and writing tasks scored on a nine-point scale had exact agreement rates of between 55% and 72%.

In Figure 15 we show the range of exact agreement rates in the literature (indicated by the dashed lines) and the exact agreement rates for 6 mark items in the 7 subjects (business studies, geography, physics, religious studies, English language, history and psychology) in England for the 5 years for which we hold seeding data.

Broadly speaking, we can see that rates of exact agreement are within the bounds suggested by the international literature, suggesting that marking consistency (as measured by exact agreement) in England is not dissimilar to that elsewhere. There are some differences between subjects, with geography having rates similar to the lowest rates seen in the literature for 6 mark items. Figure 15 also indicates that marking consistency is very stable over the time period for items with this tariff.

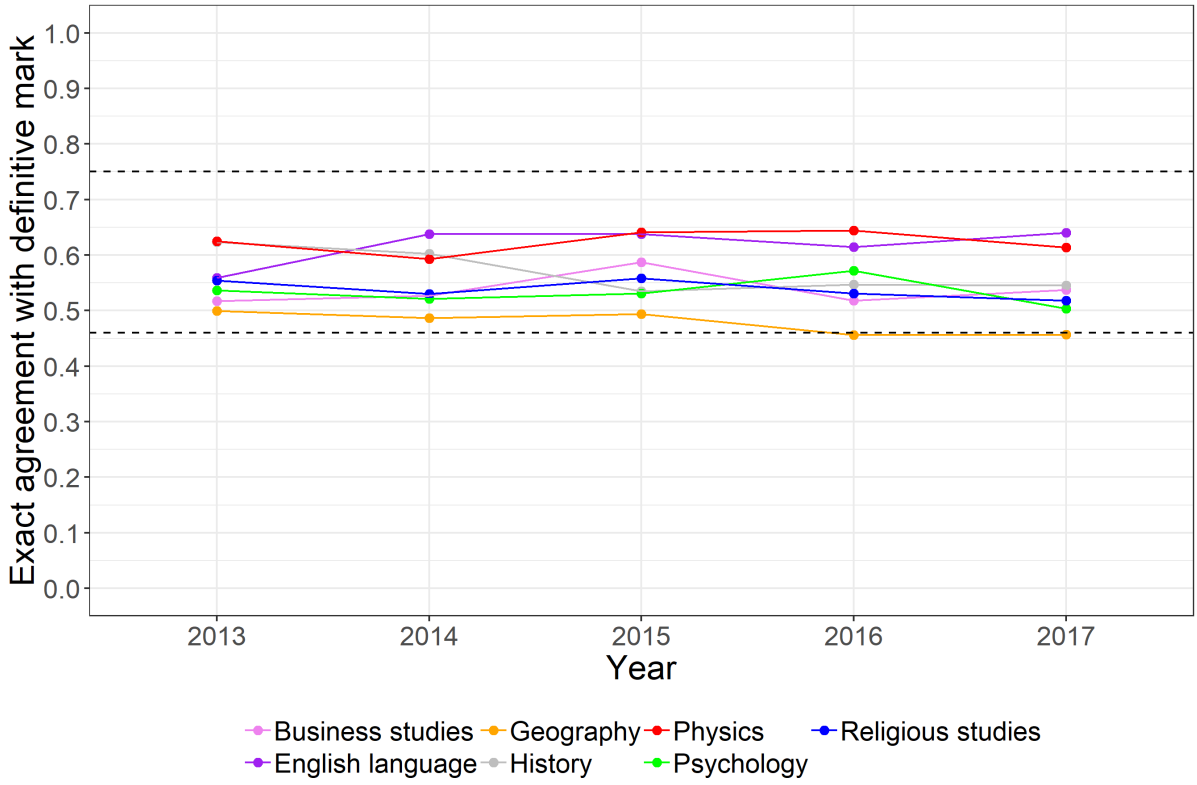


Figure 15. Exact agreement for 6 mark items, compared with international literature and over time for seven subjects.

The international literature also includes data on agreement within ± 1 mark (or 'adjacent agreement'). These rates range from 87% at the lowest to 98% at the highest. Figure 16 plots the adjacent agreement for the same 6 mark items, compared with international literature and over time for same 7 subjects. In this graph, we can see that overall these are within or above the range indicated by international literature. For both of these graphs, we are plotting the overall levels of exact/adjacent agreement for all 6 mark items within a subject. Within this, there are many 6 mark items, and there is some variability within these.

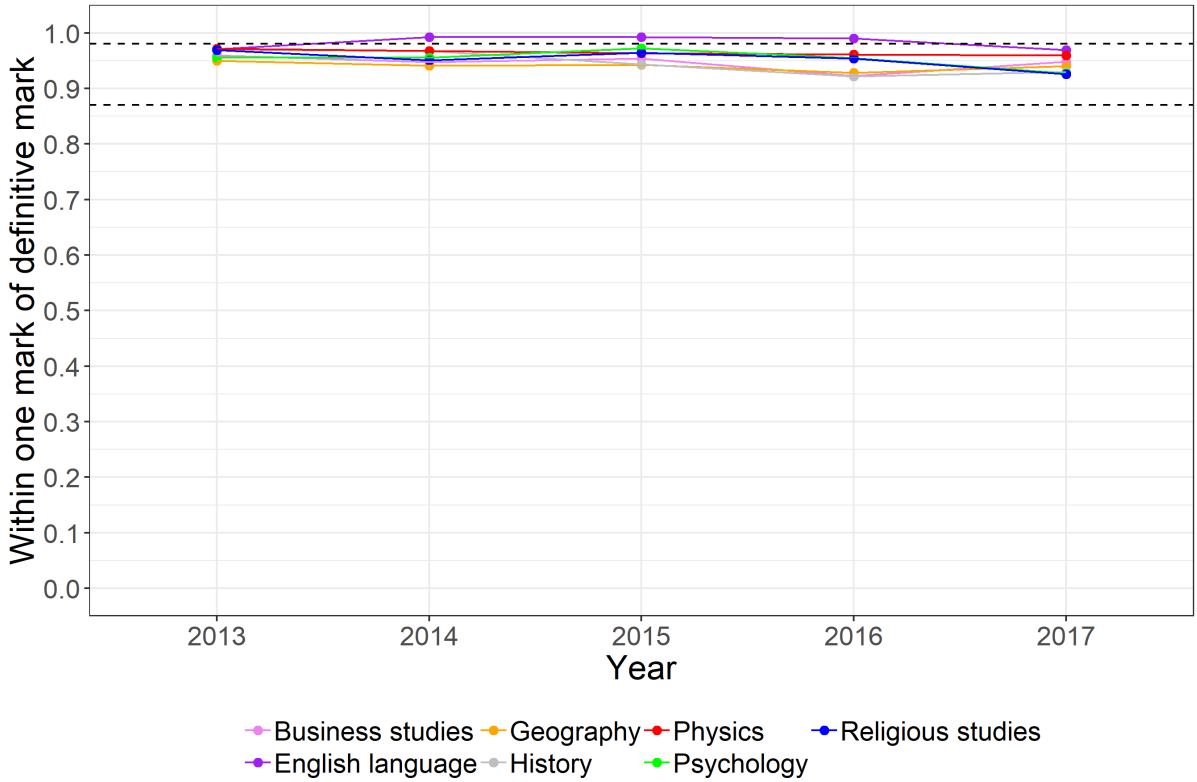


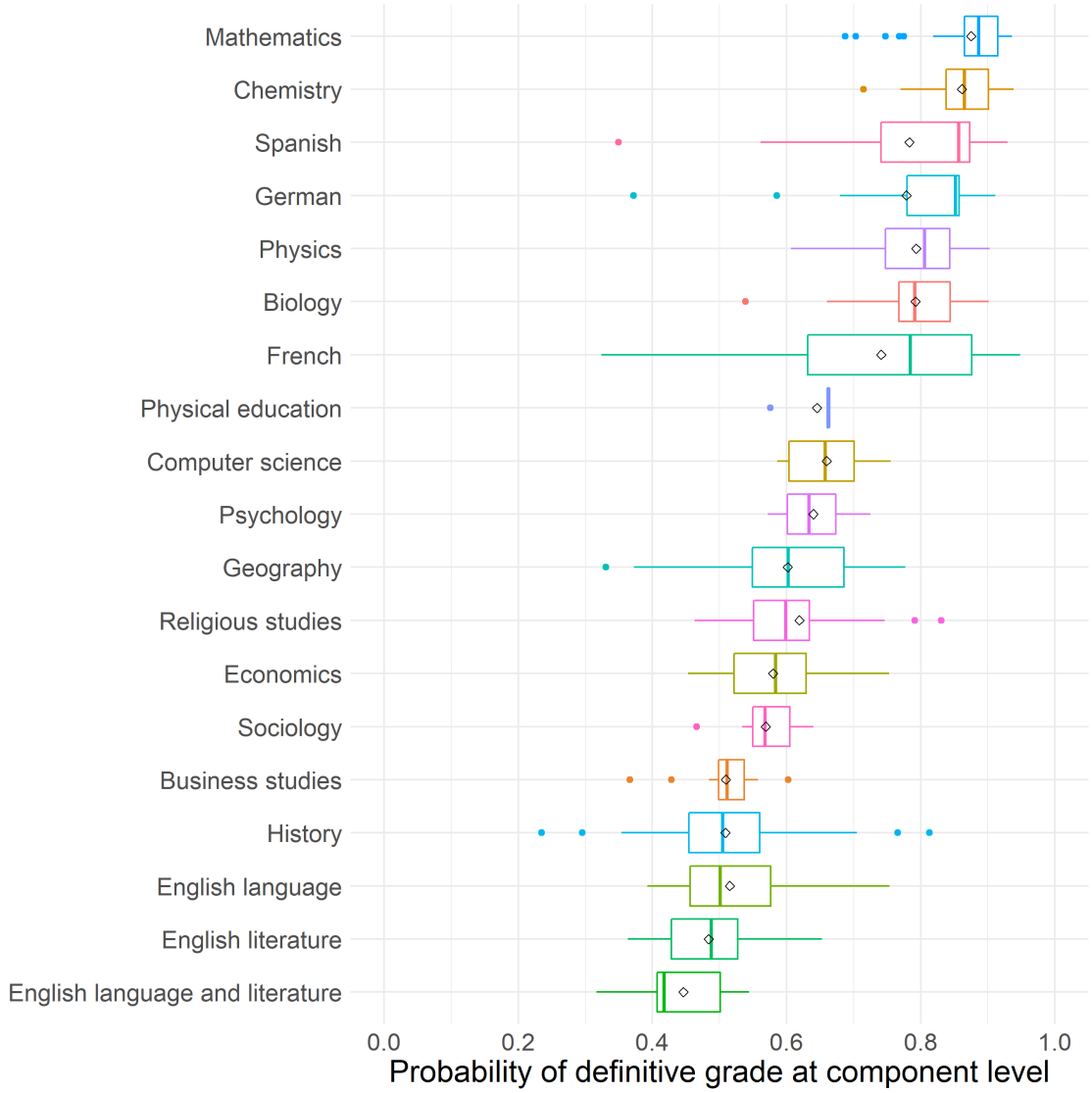
Figure 16. Adjacent agreement (within ±1 mark) for 6 mark items, compared with international literature and over time for seven subjects.

3.6 Exploring the possibility of benchmarks for marking consistency metrics

A stated aim of the metrics in Ofqual's 2014 report, was that the proposed metrics would be used to define acceptable levels of marking consistency in different assessment types. It was envisaged that such metrics may help to drive improvements in the marking consistency of general qualifications. Deciding on what defines acceptable levels of marking consistency is not straightforward. We have seen earlier (eg Figure 7, reproduced below for easy reference, and Figure 11) that there are varying levels of marking consistency between and within subjects. What appears to be very good consistency for one component in a particular subject, may not look so good when compared with components in other subjects or with

other types of item or mark scheme. For example, a long extended response question will be marked less consistently than a low tariff relatively objective question. These differing features of an assessment are in place for reasons of validity. The aim of these metrics is not to drive different styles of assessment and so it is crucial that any attempts to benchmark consistency metrics is properly contextualised. Put simply, to have one benchmark, against which all components from subjects are compared, would be a very blunt tool for comparison. It would neither set a standard for marking consistency that was achievable for many components in those subjects at the bottom of Figure 7, nor work to improve components in those subjects at the top of Figure 7.

Thus, it seems sensible to pursue some sort of means of comparison which is contextualised and involves making comparisons between components that are similar in subject and/or assessment type.



Reproduction of Figure 7 included here for easy reference: Boxplot of the probability of a candidate being awarded the definitive grade at component level, by subject. The mean probability is denoted by the black diamonds.

One possible way to do this is to identify the components with lower or the lowest consistency metrics within a subject area. Component level metrics are the chosen level of focus because (a) they are the building block of qualification level metrics and (b) because exam boards tend to standardise and monitor markers by component, so this level of focus has a practical significance.

Initially, within a subject, all components are ranked from high to low by the probability of being awarded the definitive grade within a subject. Once ranked, there are several simple approaches to decide on benchmarks, or thresholds, for acceptable consistency. The following are examples of suggestions for identifying thresholds for flagging components with potentially unacceptable levels of marking consistency, all based on the idea of ranking components within subjects or subject groups:

1. identify a simple threshold based on a fixed value below the mean probability for the subject, such as 0.05, 0.1 or 0.2 below the mean.
2. place components into equal sized categories such as quartiles, or octiles or deciles based upon their rank order. In this approach, the lowest category or group of categories would be flagged as potentially having unacceptable levels of marking consistency.
3. plot the components as boxplots and to consider any outliers as components where the marking consistency is too low (Figure 7). Outliers on the boxplots are those which are beyond 1.5 times the interquartile range from the lower quartile.
4. calculate the z-score for each component using

$$z - score = \frac{value - mean}{standard\ deviation}. \quad (3)$$

Any components with a z-score below -1.96 (the 95% confidence interval) would indicate a component with unacceptable marking consistency.

A comparison of some the methods of flagging potentially unacceptable levels of marking consistency may be seen in Table 2, with the number of components flagged in each method in Table 3. Each method has advantages and disadvantages; for example applying a simple threshold of 0.05 below the subject mean is simple but does not take into account the differing distributions within subjects; whereas using the lower quartile or z-score does take into account the differing distributions within subjects but could lead to very low thresholds (eg around 0.3 for English literature) which would be considered unacceptable from a validity perspective. A possible solution is to present a suite of benchmarks as seen in **Error! Reference source not found.** and Table 3. In future work, thresholds will be explored further. In deciding which are the most appropriate thresholds, we should take into account both the number/proportion of components flagged, as well as the public acceptability of the threshold. It is also possible that grouping by subject does not provide appropriate contextualisation in some cases (eg where there are assessments of very different character and associated marking consistency).

Table 2 Comparison of different thresholds of marking consistency.

Subject	0.05 below subject mean	Lower quartile	>1.5*IQR below lower quartile	$z < -1.96$
Biology	0.742	0.766	0.655	0.650
Business studies	0.459	0.491	0.448	0.401
Chemistry	0.811	0.838	0.744	0.766
Computer science	0.610	0.587	0.506	0.527
Economics	0.530	0.518	0.364	0.433
English language	0.465	0.413	0.280	0.341
English language and literature	0.396	0.400	0.273	0.311
English literature	0.460	0.432	0.264	0.313
French	0.691	0.602	0.353	0.396
Geography	0.552	0.545	0.347	0.379
German	0.728	0.680	0.697	0.470
History	0.459	0.454	0.296	0.321
Mathematics	0.825	0.864	0.792	0.770
Physical education	0.595	0.661	0.659	0.569
Physics	0.755	0.754	0.610	0.688
Psychology	0.590	0.582	0.513	0.547
Religious studies	0.569	0.551	0.447	0.403
Sociology	0.519	0.534	0.483	0.456
Spanish	0.733	0.736	0.558	0.462

Table 3. Number of components flagged on the basis of four different thresholds.

Subject	Total number of components	0.05 below subject mean	Lower quartile	>1.5*IQR below lower quartile	$z < -1.96$
Biology	39	7	10	1	1
Business studies	15	2	4	2	1
Chemistry	37	5	10	1	1
Computer science	6	2	2	0	0
Economics	19	8	5	0	0
English language	19	7	5	0	0
English language and literature	11	2	3	0	0
English literature	27	8	7	0	0
French	14	4	4	0	1
Geography	31	8	8	1	2
German	12	3	3	2	1
History	64	17	16	2	2
Mathematics	60	8	15	6	5
Physical education	5	1	2	1	0
Physics	45	12	12	0	0
Psychology	15	4	4	0	0
Religious studies	13	4	4	0	0
Sociology	7	1	2	0	0
Spanish	14	3	4	1	1
Percentage flagged		23.4%	26.5%	3.8%	3.3%

3.6 Optionality

In the 2016 metrics report, a potential limitation of the metrics was the assumption needed to evaluate the consistency of marking in cases where an assessment included optionality. Optionality means that some or all of the questions are optional. For example, candidates might have to answer all questions in section A, and choose one question from section B.

Optionality might pose an issue for marking consistency metrics at component or qualification level if there are differences in marking consistency between optional items. Such differences may arise for a number of reasons, including the selection of seeds for a particular question, or be a function of the question or the allocation to a (subset) of examiners. Nevertheless, checks on the marking of individual items are straightforward, as long as each optional item is represented in the seed or double marking process. Component level comparison of optional routes, though, is more complex. It requires assimilation of the cumulative effects of optionality across the

whole assessment. We commissioned an external consultant to look into this using a mixture of actual seeding data (fully anonymised) and simulations based on actual data. The report is provided in the Appendix.

For all analysis presented so far, in order to create component or qualification level metrics, optional questions in a component have been collapsed into a single item. Where the rubric required a candidate to choose one option from many, and where the breakdown of that question (in terms of mark allocation) was consistent between options, it was assumed that each optional question had the same mark-remark reliability. There is some evidence to suggest that this assumption may not always be valid. Figure 17 shows the mean difference between the seed and examiner marks for optional items within sections A and B of a legacy component. The 95% confidence intervals surrounding the mean difference suggests that there are some differences in the consistency of marking across the component. For example, the marking of the fifth item in section A appears lenient in comparison with that of the first item.

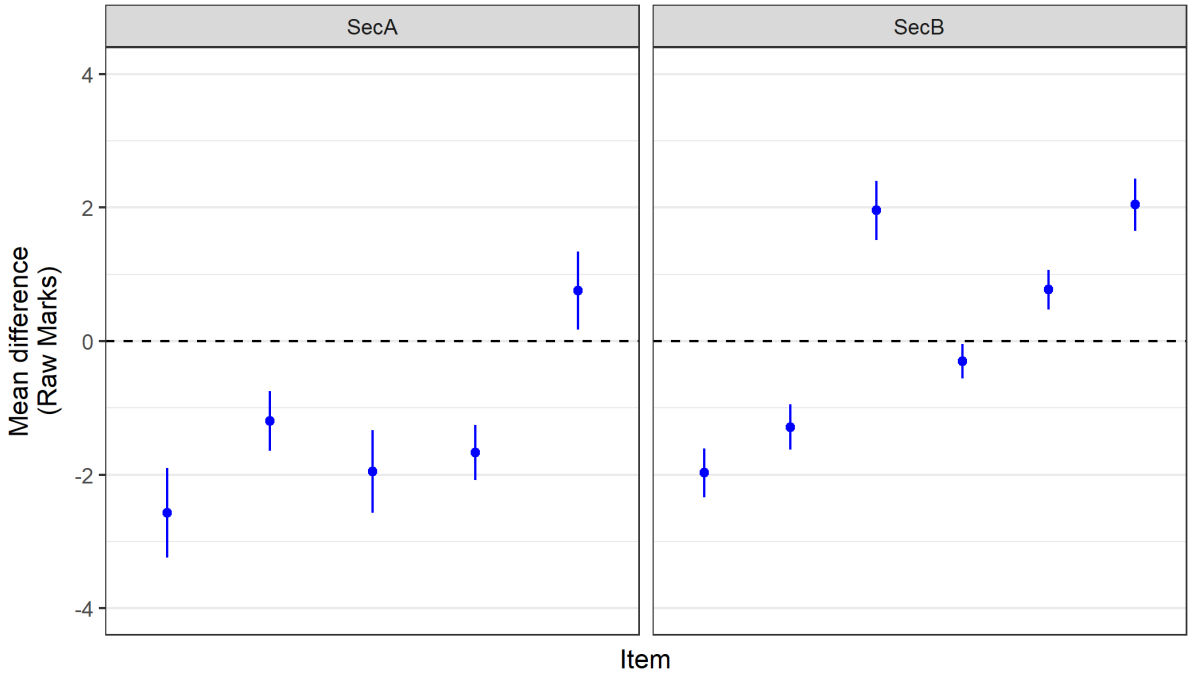


Figure 17. Mean difference with the definitive mark and the 95% confidence interval for the optional items within sections A and B for the legacy component.

Sitting beneath the item level statistics is the sampling process from which the data is derived. There are more than 20 questions in this particular component but seed data is only available for 11 questions. While it would be desirable to include a range of responses in the seeding process to ensure that all aspects of the mark scheme are effectively applied, the seeding process is primarily designed to check the marking of examiners. Creating marking consistency metrics is secondary. However, the small number of seed items for optional questions suggests that inferences based on these metrics might lack rigour.

By using pseudo-candidates and multilevel modelling as outlined in the appendix it is possible to model different routes through the component. Instead of combining optional questions, each item is treated individually; each item has its own unique

estimate of the level of marking consistency estimated from the difference from the definitive mark. The probability of being awarded the definitive grade is then calculated by weighting the optional questions according to the number of responses in the population. While this creates a single statistic for each component, it is also possible to generate metrics for each of the optional routes available within the component (as well as within the qualification).

In this case study we looked at a 15 different legacy GCSE and A level components. The rubric for these components was straightforward; in any given section with optionality, candidates chose one question. It is worth noting that more complex rubrics will require further investigation. However, this case study provides a first step in evaluating the effect of optionality on marking consistency metrics.

The steps to calculate the probability of the definitive grade are set out in the appendix. Figure 18 shows the probability of being awarded the definitive grade using the original method and the revised method where each optional question is treated individually. As the weights for each optional question are not available the revised method is repeated five times with different random weightings of optional questions. Each line in Figure 18 is surrounded by a 95% confidence interval denoted by the shaded area (the 'ribbon'). It is observed that the probability of the definitive grade varies between 0.4 and 0.7. The probability of achieving the definitive grade is not significantly different between the original and revised methods nor is the rank order of components apart from component 7. It appears that the modelling of optional questions has little influence on the overall conclusions that would be drawn regarding marking consistency for GCSE and A level components in this simulation. Interestingly, it is observed that the component in which the different optional weights appears to have the greatest impact on the metric is the component which has one of the lowest levels of optionality (see component 7 in Figure 18). In this particular component, there are several compulsory questions but only two optional questions. The first optional question was marked on average 2.5 marks more severely than the definitive mark whereas the second question was marked on average 1.2 marks more leniently than the definitive mark. Perhaps unsurprisingly, when there are large differences in marking consistency between options and there are only few options, the component level metric is sensitive to the weight of each optional question.

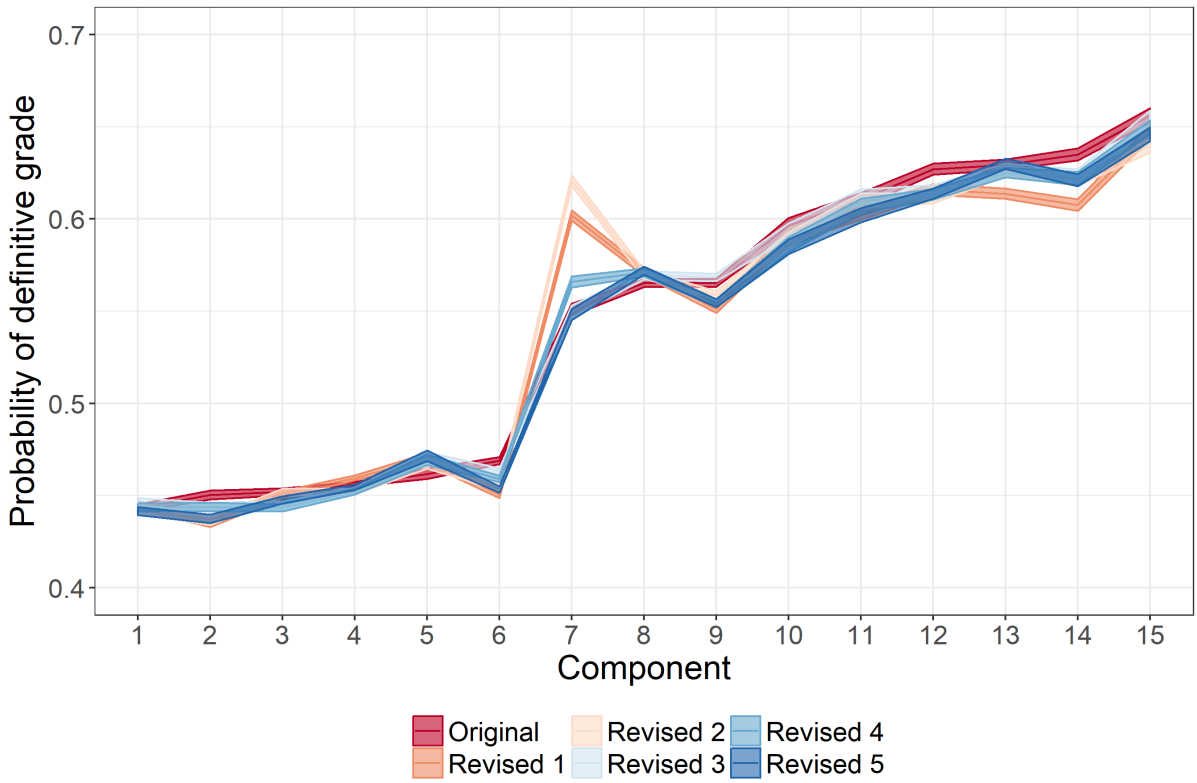


Figure 18. Probability of being awarded the definitive grade (line) and the 95% confidence interval (ribbon) evaluated using the original and revised metric calculation.

It is also possible to present route level metrics within a component. For component 2 in Figure 18 there are 30 possible routes to a component grade[12]. The probability of achieving the definitive grade at component level for each route is shown in Figure 19. Unlike Figure 18, where five different weights were used to illustrate the sensitivity of the model, the route level metrics are represented with one set of weights. The confidence intervals around each route differ in width and illustrate the fact that some estimates are based on very few data points. When the simulation is run to create 5,000 pseudo candidates, only 10 are entered for route 1 compared to nearly 400 for route 11.

[12] Had all optional questions been represented in the seeding data, there would have been well over 150 routes to a component grade.

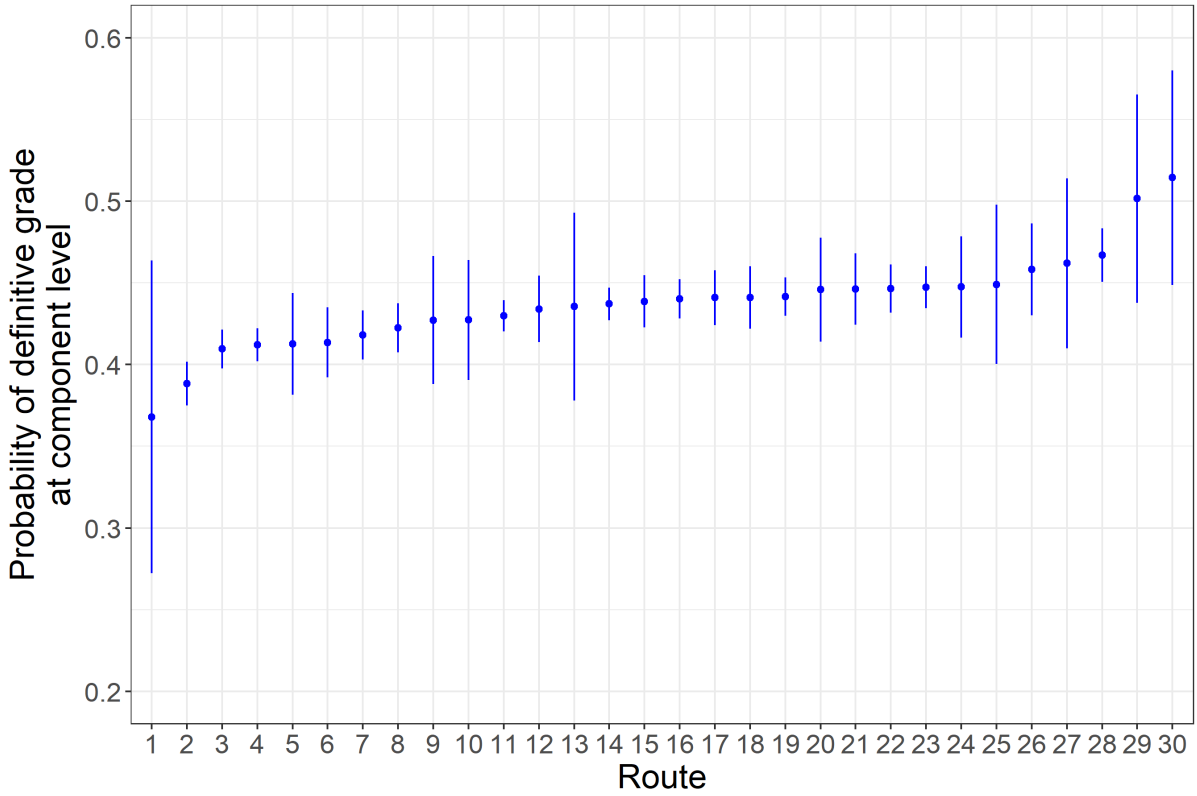


Figure 19. Probability of the definitive grade (solid circle) and associated 95% confidence interval for the 30 different routes in component 2.

Perhaps a more informative approach to view route level metrics is as a measure of potential variation in marking consistency across routes. For component 2, the mean consistency across routes is 0.46 and standard deviation of marking consistency across routes is 0.03. The marking consistency could be reported as 0.46 ± 0.03.

4. Conclusions

A series of metrics are presented in this report, as are the assumptions and conditions necessary to derive them. We have also reported on qualification level metrics for the first time, made possible since the introduction of fully linear qualifications and the removal of internal assessment from a number of subjects. It has been observed that qualification level metrics are generally higher than the components from which it is comprised.

Consideration of the practical uses of metrics, such as the derivation of thresholds to identify acceptable and unacceptable levels of marking consistency, have been discussed. In deciding which are the most appropriate thresholds, we should take into account both the number and/or proportion of components flagged, as well as the public acceptability of the threshold.

This report also shares some data (based on the marking of 6 mark items) comparing the marking in England with marking elsewhere. This provides some indication that the marking in England is of similar levels of consistency to elsewhere. This report, using the same data, also indicates that marking consistency

over time (between 2013 and 2017) appears to be relatively stable; it has neither deteriorated nor improved.

There is a range of values of metrics reported here, with some definite subject patterns. The median probability of receiving the definitive qualification grade varies by qualification and subject, from 0.52 (English language and literature) to 0.96 (Mathematics). The probability of receiving the definitive grade or adjacent grade is above 0.95 for all qualifications, with many at 100%. This is not to say that there are not components or qualifications where the marking consistency cannot be improved. Through identifying appropriate thresholds of acceptability, exam boards should channel additional resource and support to those components or qualifications which most need improving.

All future work with metrics needs to proceed with some caution in order to manage the risk that any use of thresholds or benchmarks does not compromise the live on-line monitoring procedures and hence the actual quality of marking which is the very thing we wish to improve.

5. References

Bramley, T., & Dhawan, V. (2010). Estimates of Reliability of Qualifications. Published in Eds Opposs, D., and He, Q., (2012) The Reliability Compendium, Ofqual, Coventry.

Brown, G., Glasswell, K. & Harland, D. (2004). Accuracy in the scoring of writing: Studies of writing assessment system. Assessing Writing 9, 105-121.

Cresswell M., (1986). Examination Grades: how many should there be? British Educational Research Journal 12, 37-54.

LeMahieu, P., Gitomer, D., & Eresh, J. (1995). Portfolio in large-scale assessment: Difficult but not impossible. Journal of Educational Measurement: Issues and Practice 14, 11–28.

Longford, N. (1996). Adjustment for reader rating behavior in the Test of Written English. ETS Research Report RR-95-39, TOEFL-RR-55.

Ofqual. (2014). Review of quality of marking in exams in A levels , GCSEs and other academic qualifications. Ofqual, Coventry.

Penny, J., Johnson, R., Gordon, B. (2000) The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. Assessing Writing 7:2 143-164

Rhead, S., Black, B., & Pinot de Moira, A. (2016). Marking consistency metrics. Ofqual, Coventry.

Stockford, I., & He, Q. (2014). Reporting of assessment functioning statistics for regulated qualifications - a paper for discussion.

Supovitz, J. A., MacGowan, A., III & Slattery, J. (1997). Assessing agreement: Interrater reliability of portfolio assessment in Rochester, New York. Educational Assessment 4, 237–259

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). A review of literature on marking reliability research.

Appendix A – Dealing with optionality in marking consistency metrics

Report by Anne Pinot de Moira, March 2018

1. Introduction

In November 2016, Ofqual published a report on marking consistency (Rhead, Black, & Pinot de Moira, 2016). It proposed a series of metrics from which to evaluate the consistency of marking in GCSE and A level components. Listed among the potential limitations of these metrics was the assumption needed to evaluate the consistency of marking in cases where an assessment includes optionality. This paper explores the problems inherent in the calculation of marking metrics where optionality exists.

The analysis builds upon previous work, using the current metrics as the starting point. It is presented in the form of a case study using GCSE and A-level components offered in summer 2016, all containing optional routes.

2. Data

The data include seed mark information for fourteen components and, for one further component, double marked information. All the components have optional questions but, in each case, the rubric is reasonably straightforward. Candidates are required to choose one from a number of questions within a section.

3. Marking consistency metrics

Current metrics

The main body of the current report provides an update of marking consistency metrics for a range of GCSE and A level qualifications (Rhead, Black and Pinot de Moira 2016). It excludes qualifications which are not marked on-screen. The data are presented at an item and component level. Due to the difficulties with assessing marking consistency for internally-assessed and -marked components, no qualification level metrics are yet available for those qualifications containing non-examined assessment.

The original solution to optionality

Optionality is only really an issue if there are differences in marking consistency between optional items. These differences may arise from ambiguities with the question, responses, mark scheme or marking. Alternatively they may be a function of selective allocation of examiners to items. Nevertheless, checks on the marking of individual items are straightforward as long as each optional item is represented in the seed or double marking process; more of which later. Component-level

comparison of optional routes, though, is more complex. It requires assimilation of the cumulative effects of optionality across the whole assessment.

In the original report on marking consistency, to create a component-level metric, Rhead et al (2016) treated optional questions as a single item. Where the rubric required a candidate to choose one option from many, and where the breakdown of that question (in terms of mark allocation) was consistent between options, it was assumed that each optional question had the same mark-remark reliability.

The issue of optionality, and the problem of understanding its effect on the functioning of an assessment, was also discussed by Stockford and He (2014). They, too, noted that the options could be treated as a single entity. However, they favoured the proposal that, for components with optional questions, each route through the qualification should be analysed separately. In other words, each possible route should be treated as a separate component. This proposal was not pursued because complex optionality could give rise to many different routes through the qualification[13].

In the original report, Rhead et al (2016) described two component-level metrics. The first based on the sum of independent variables (CL1, section 4.2, page 15). This metric was favoured because of its transparency and because it required fewer assumptions; the drawback being that it effectively precluded the modelling of optionality. The second metric was based on pseudo-candidates (CL2, section 4.2, page 17). It required more assumptions but allowed flexibility to model different routes through the assessment.

The proposed revision

Revisiting the issues surrounding optionality allows a fresh look at the problem.

Item-level marking consistency statistics provide an operational view of any areas where the system might show opportunities for improvement. They allow scrutiny of individual questions which have given rise to inconsistent marking. Where all items are compulsory, the value of this information is in the feedback loop it creates for assessment writers as well as those who lead marking training. Where items are optional, item-level statistics highlight potential areas of inequity in the assessment and may raise questions about the training or standardisation of examiners. Further investigation would be needed to understand whether any inequity arises from the assessment, the examiners or both.

Component-level statistics allow comparisons within and between subjects. Potential revision of the component-level statistics to adjust for optionality focuses on reformulating the probability of being awarded the definitive grade, the second component-level metric described above (CL2). Instead of combining optional questions, each item is treated individually. Therefore, each item has its own unique estimate of the level of marking consistency estimated from the mark-remark quality control process. The probability of being awarded the definitive grade is then calculated by weighting the optional questions according to the number of responses in the population. While this creates a single statistic for each

[13] In fact, Stockford and He (2014) also recognised this as a problem and suggested that data for the most popular 10 routes should be reported.

component, it is also possible to generate metrics for each of the optional routes available within the component (as well as within the qualification). That said, for many components the number of routes could be enormous and, in operational terms, it is unclear the additional value of these route-level statistics beyond that of the item-level statistics.

The rubric for the GCSE and A level qualifications included in the case study is straightforward: in any given section with optionality, candidates simply had to choose one question. Computationally, this rubric is also reasonably straightforward. More complex rubrics would require more complex programming. So this study provides a first step in evaluating the effect of optionality on marking consistency metrics.

4. Results

Item-level statistics

Figure A1 shows the mean difference between the seed and examiner mark for optional items within sections A and B of a single component. The confidence intervals surrounding the mean estimates suggest that there are some differences in the consistency of marking across the options. For example, the marking of item 5 appears lenient in comparison with that of 1.

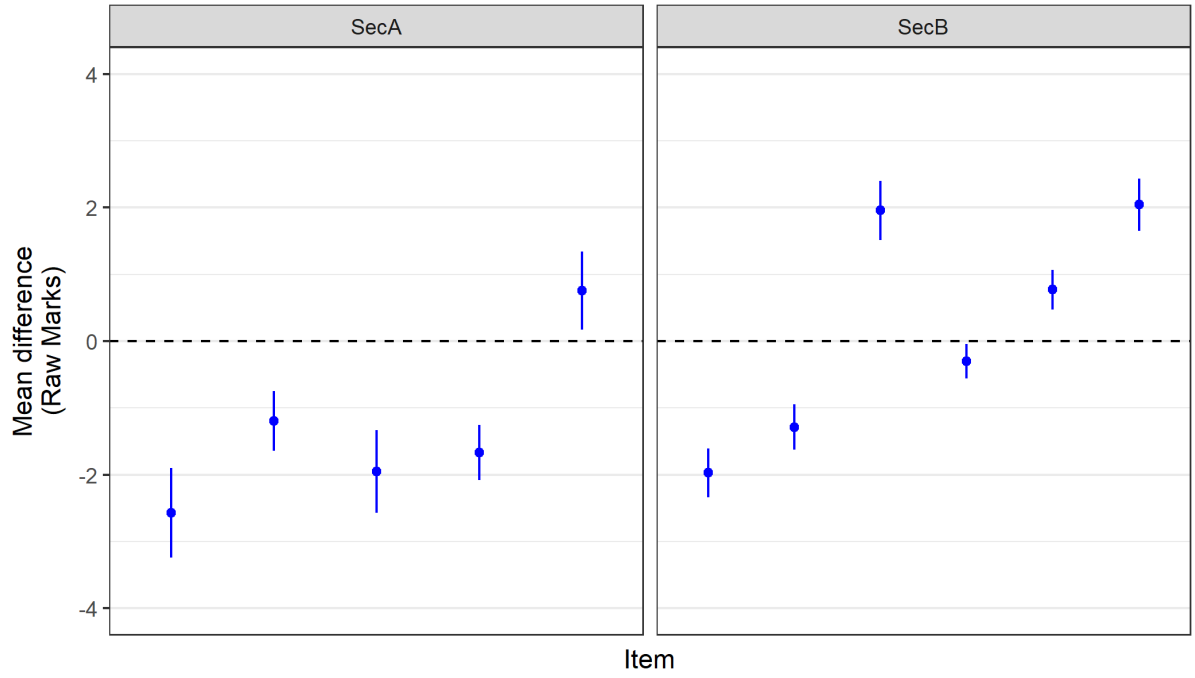


Figure A1 - Mean mark-remark difference and 95% confidence interval for the optional items within sections A and B of one component.

Sitting beneath the item level statistics, however, is the sampling process from which the data are derived. First, there are no seeds for over half of the optional items on the paper. Second, where there are seeds, even if the number of seed

remarks is relatively high, the number of unique seeds is sometimes low and may not represent the number of responses overall.

As a percentage of all responses the number of seeds is extremely low. While it would be desirable to include a range of responses in the seeding process to ensure that all aspects of the mark scheme are effectively applied, the seeding process is primarily designed to check the marking of examiners. Creating marking consistency metrics is secondary. However, the paucity of seed items for optional questions suggests that inferences based on these secondary metrics might lack rigour.

A comparison of the original and revised component-level metric

The steps to calculate the probability of the definitive grade are set out in Rhead et al (2016). As previously discussed, the original component-level metrics collapse optional questions to a single item; assuming they all have the same mark-remark consistency. In contrast, for the revised metric, the mark-remark difference is evaluated separately for each optional question[14]. Ideally, pseudo-candidates would be generated such that the optional questions are weighted in the pseudo-population in proportion to the true population. The true population weights would be derived from the response frequency for each question. However, because these population weights are not available for this investigation, five sets of randomly generated weights are used to illustrate the extent to which optionality might influence the component-level marking consistency metric.

Figure A2 shows, for each component, the estimated probability of gaining the definitive grade using the original and revised metric calculation. The red line represents the original calculation where no account was taken of optionality. The remaining lines represent the effect of optionality using the five sets of random weights. Each line is surrounded by 95% confidence intervals denoted by the shaded area. The y-axis has been contracted to emphasise the differences as all probabilities lie between 0.4 and 0.7.

The two calculations do give rise to different statistics but, on the whole, these statistics are not significantly different nor do they change the rank order of the components in terms of consistency. So, the modelling of optional questions has little influence on the overall conclusions that would be drawn regarding marking consistency in these components. It is interesting, however, that the component in which the different optional weights appears to have the greatest impact is the component with the lowest level of optionality. Component 7 is a paper in two sections. Section A is compulsory and, for section B, the candidate must answer one from two questions. In the summer 2016 examination, the first question in section B was marked on average 2.5 marks more severely than the seed. In contrast, the second question in section B was marked on average 1.2 marks more leniently. Perhaps unsurprisingly, when there are large differences in marking consistency between options and few options, the component-level metric is sensitive to the weight of each optional question.

[14] But only when there is seed or double marked data.

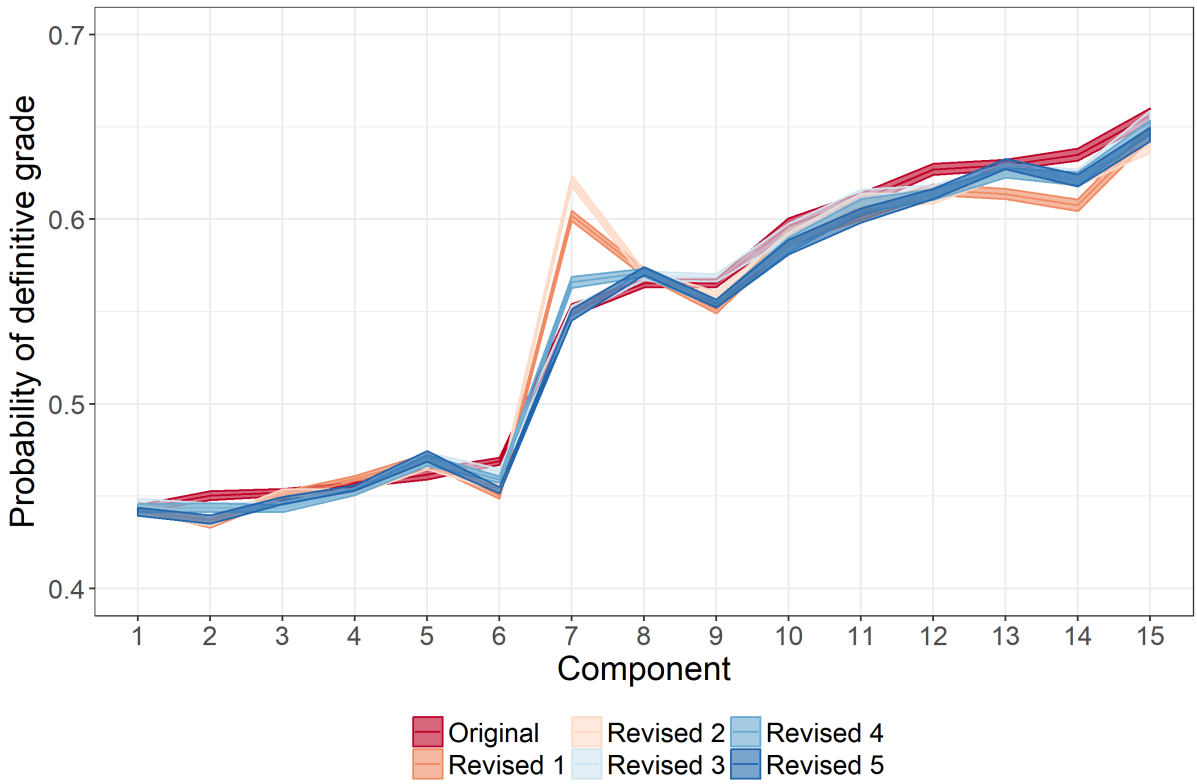


Figure A2 - Overall probability of the definitive grade (and associated 95% confidence interval) evaluated using the original and revised metric calculation

A route-level metric

It is also possible to present route-level metrics within a component. For component 2 in Figure A2 there are 30 possible routes to a component grade[15]. The probability of achieving the definitive grade at component level for each route is shown in Figure A3. Unlike Figure A2, where five different weights were used to illustrate the sensitivity of the model, the route level metrics are represented with one set of weights. The confidence intervals around each route differ in width and illustrate the fact that some estimates are based on very few data points. When the simulation is run to create 5,000 pseudo candidates, only 10 are entered for route 1 compared to nearly 400 for route 11. Adjusting the algorithm to create 5,000 pseudo candidates per route would vastly increase the processing time even with the simplest of

[15] Had all optional questions been represented in the seeding data, there would have been well over 150 routes to a component grade.

optional structures.

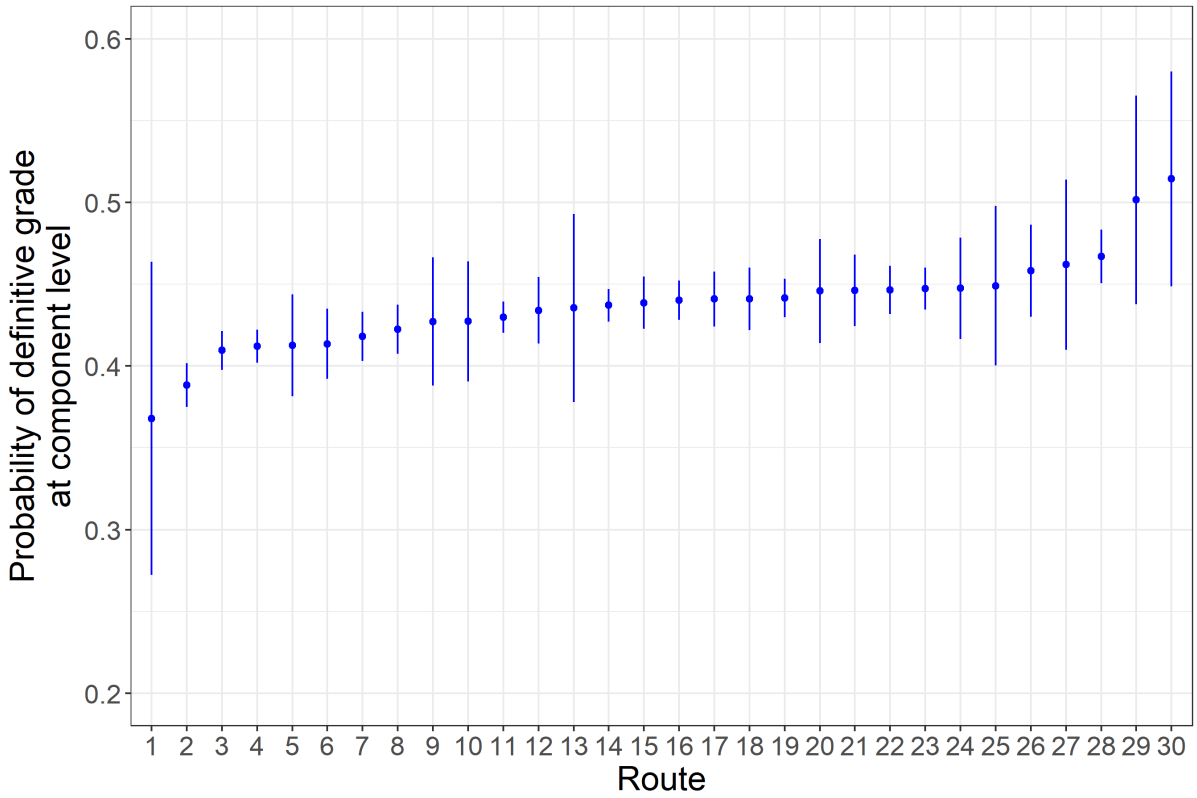


Figure A3. Probability of the definitive grade (and associated 95% confidence interval) for the different routes in component 2.

Perhaps a more informative way to view the route-level data is as a measure of the potential variation in marking consistency across routes. For component 2, the standard deviation of marking consistency across routes is 0.028. This might more usefully be expressed as a coefficient of variation so that comparisons could be made between components. The coefficient of variation for 2 is 0.066 and for 1, which is identically structured, is 0.052. Whether this difference is acceptable given the sensitivity of the model would require further investigation.

Evaluation

While the revised component-level marking consistency metric might represent a step forward in the evaluation of marking in GCSE and A level qualifications, it is still based on many assumptions and limitations. Some of these assumptions can be tested but some cannot.

Missing data

Not all optional questions are selected as part of the quality control process. While double-marked items are often allocated at random to examiners, seed items are mostly selected in advance of the marking period on the basis of their

characteristics and, perhaps more importantly, their availability[16]. Less popular questions are, therefore, less likely to be represented.

For the primary purpose of quality control, it is a moot point whether checks should be made on all optional questions regardless of uptake. On the other hand, for the ancillary purpose of assessing the impact of optionality on marking consistency, mark-remark data must be available for all options. Without data for each option, there is no way of comparing optional marking consistency without imputing data. Any imputation method would require some assumption about the relationship between options; thereby defeating the object of comparing the options.

If the effect of optionality on marking consistency is to be modelled, it should be modelled only where data are available. Alternatively, guidance surrounding the quality control process should be strengthened to require that all optional questions are included at least once in the seeding or double marking process.

Sensitivity

To date, little has been done to test the sensitivity of the component-level marking consistency metric regardless of whether optionality is explicitly modelled. Findings from the case study suggest that the model is not very sensitive to optionality but our understanding is limited to the simple examples of optionality observed in the legacy qualifications in this simulation. In many qualifications, the rubric is more complex and the difference between the requirements of optional questions is greater. The legacy assessments in this study are often served by generic mark schemes designed to unify understanding of performance across the options.

Furthermore the marking consistency metrics are based on an underlying multilevel model and artificially created pseudo-candidates. The multilevel model is simple and takes no account of potential non-linear relationships between mark-remark difference and features of the assessment. The pseudo-candidates are created with a fixed correlation between items. In all previous analysis this has been set to about 0.4. To contextualise the findings presented in Figure A2, Figure A4 shows the effect of adjusting the between-item correlation.

Generally, the higher the correlation between item marks, the higher the probability of being awarded the definitive grade. The level of variation between estimates of the component-level marking consistency metric is similar to that seen in Figure A2. The model is therefore sensitive to the parameters used in the estimation.

Perhaps more intriguingly, Figure A4 reveals that the effect of varying the between-item correlation differs considerably between components. For component 5, the estimate of marking consistency ranges between 0.44 and 0.63. This component differs from the others in that it has vastly more individually recorded item marks; 62 in total compared with fewer than 15 for each of the remainder. This suggests that the model might also be sensitive to the number of items on the paper.

[16] There is a small window of time between the date of the examination and the start of the marking period when scripts are being scanned and uploaded to the system. Seed items are also selected during this window and choice is therefore limited because of the concurrent nature of the two activities. Further details of the quality control processes across the awarding organisations are given in Rhead et al (2016).

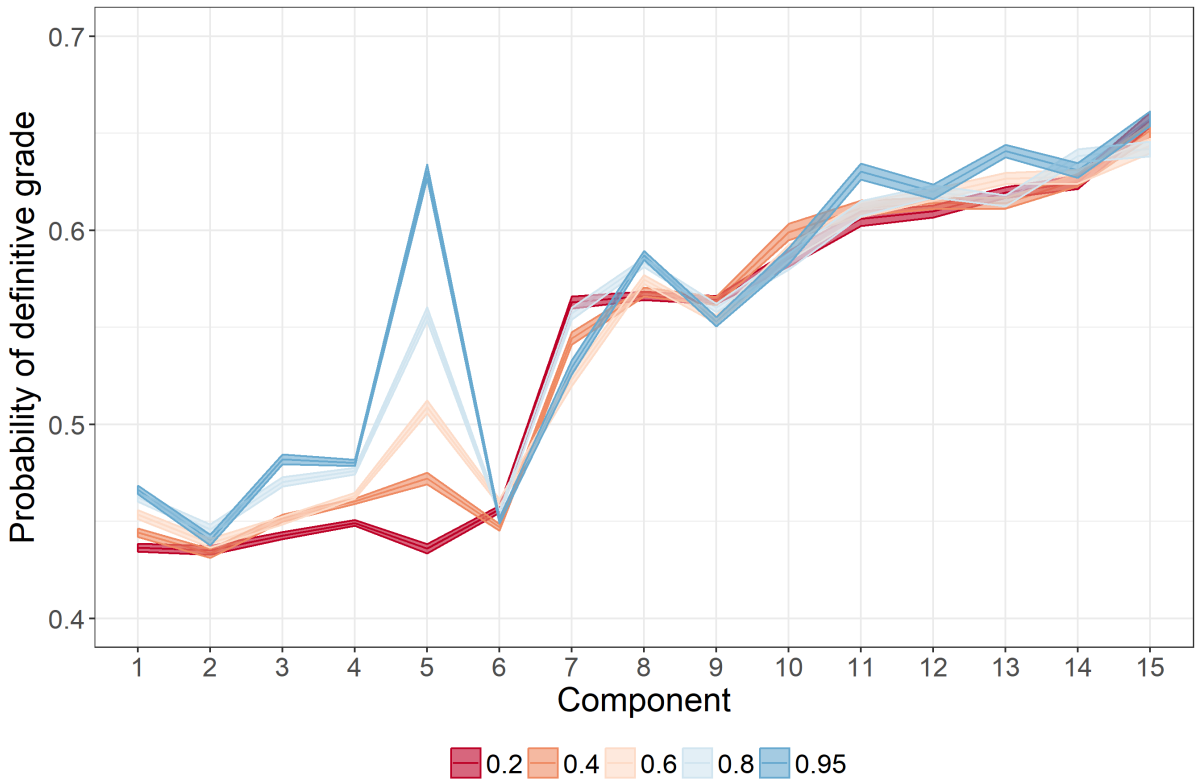


Figure A4: Overall probability of the definitive grade calculated using the revised metric and altering the between-unit correlation

It should, however, be remembered that a mark scheme and seeding mechanisms are set up firstly for the purpose of quality assurance. In section A of component 5, candidates are required to answer one question from five. Each question has two parts. The first part is worth 20 marks and is judged as a single entity. In other words, one mark is recorded on the marking database. The second part is worth 40 marks and is judged separately for each of four assessment objectives that it tests. In total, the single question in section A gives rise to five discrete marks and there are mark-remark differences for each one. To make sense of a candidate's response, all five parts must be marked by the same examiner. From a quality assurance perspective, it is of great importance to ascertain whether an examiner clearly understands all aspects of the marking process. For this reason, the practice of atomising a response might be regarded as effective. From a marking consistency perspective, such atomisation confuses the estimation of metrics particularly in situations where the question is marked by a single examiner.

Clipping

Where items are grouped together to be marked by a single examiner they are often termed as clipped. The marking consistency algorithm, CL2, is blind to clipped items, as incidentally is CL1. It processes each item separately, and assumes that they will be distributed randomly among examiners. Plainly, as seen in the case of component 5, this is not always the case. The marking is inter-related and therefore the error may be inter-related. The judgements are made about the unified response; a candidate's single thought process.

Therefore, it seems reasonable that items clipped for marking should also be clipped in the evaluation of marking consistency. A pseudo-candidate should not be modelled to get AO1 marks from question 1 and AO2 marks from question 2. The effect of clipping, largely overlooked to date, is likely to improve the estimates of marking consistency for two reasons. The first is that mark-remark differences attributed to a clipped question will be true rather than estimated. The second is that by combining the item marks within a clipped question, some degree of error cancellation will occur; albeit less than would be the case if different examiners had marked the items.

The effect of clipping related items for component 5 is to reduce the number of individually modelled questions from 62 to 13. The multilevel model fitted to the data (to attach the random marking errors to the pseudo-candidates) estimates a lower mean mark-remark difference for the question when the items are clipped than when they are unclipped. This in turn impacts upon the estimate of marking consistency as demonstrated in Figure A5.

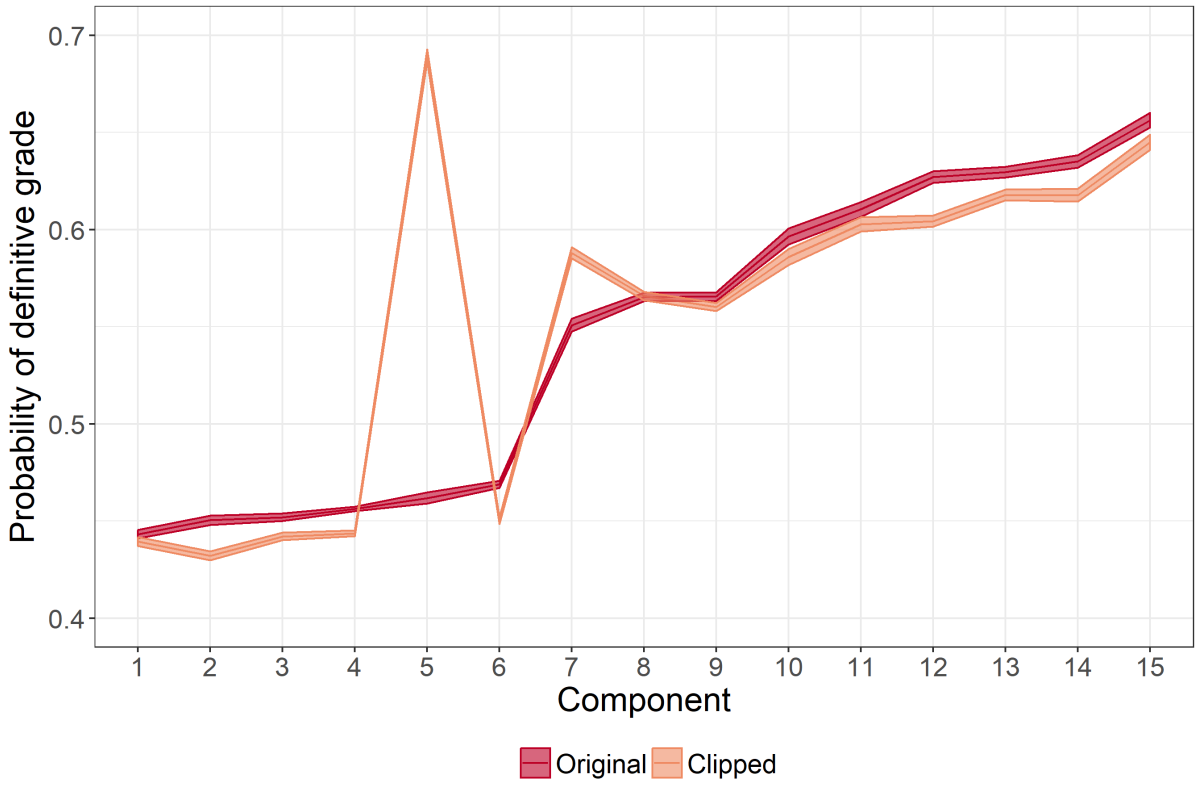


Figure A5. The impact of clipping items in component 5 on the estimate of the probability of gaining the definitive grade (and associated 95% confidence interval)

The impact of clipping, therefore, seems to be at least as significant as the impact of modelling optionality and as the impact of altering the correlation parameters. Scrutiny of the other assessments included in the case study suggests that at least four more components should have had items clipped. Indeed, in cases where whole scripts are seed marked, there is an argument that no modelling should be done but that the component-level consistency metrics should be estimated from the whole-paper mark-remark data.

Discussion

While this paper sets out to explore consistency of marking in cases where an assessment includes optionality, in so doing it reveals the sensitivity of current metrics to other influences. Some of these influences derive from assessment design and the quality control process but some from the assumptions implicit in the calculations.

Item-level metrics are largely immune to any assumptions, although comparisons between optional questions are limited by whether or not all options are included in the quality control process. These statistics may be of value in providing an operational view of areas in which at least one element of the assessment is failing.

On the other hand, component- and route-level marking consistency metrics need to be interpreted with caution. In order to evaluate marking consistency where optionality exists, it is necessary to fit a model. Changes to the parameters in this model affect the conclusions to be drawn, sometimes to a greater extent than modelling the effect of optionality itself. Significantly, the way in which items are clipped influences the value of the compound metric. Furthermore, all results are based on simulations that are initiated using randomly generated numbers. While initial values can be set as constant to ensure results are reproducible, a change to these values would lead to a change in the metrics. This is illustrated in Figure A2 with respect to the weights, which would actually be based on real data should the metric be adopted, but would also affect the generation of pseudo-candidates in perpetuity. To mitigate this sensitivity, confidence intervals have been attached to all metrics presented.

The sensitivity of the models means that, even with confidence intervals, there is a need to consider how the resultant metrics should be presented to a wider audience. Headline grabbing component-level metrics have been shown to be influenced by the design of the qualification, by the correlation between items and by the structure of marking. How should these influences be conveyed alongside the metrics? It is clear that some features of an assessment are in place for reasons of validity. In such cases, changes based on marking consistency metrics would be undesirable. In other cases, it may just be that the subject content is inherently more straightforward to mark.

For operational purposes, route-level metrics as suggested by Stockford and He (2014) seem to add little extra value over item-level metrics. It is just conceivable that they might be used as evidence for post-result enquiries but, once again, given the paucity and sensitivity of data on which they are based, they could only be used with caution.

Recommendations

Any statistic is only as good as the data upon which it is based and is only of use insofar as it is understood within context. While item-level marking consistency metrics allow for straightforward interpretation, they are still limited by the responses included in the remarking process. To improve these metrics, further research could be conducted into the seed and double mark selection process.

The coverage of seeding items/double marking items across optional questions is also worth considering further. Taking the component discussed earlier, only a third of optional items are represented in the quality control process and, even when they are represented, there is often only one response. It is difficult to know to what extent this might be found beyond the 15 legacy components used in this modelling. It might be helpful to have an inventory of optional item representation within the seeding/double marking process, and understand the extent to which this is representative of uptake of different options and routes. This might help explore the potential risks involved with including so few responses in the monitoring of quality of marking, and the challenges this provides for the evaluation of marking consistency.

Naturally, any shortcomings of the item-level metrics also become shortcomings of the component- and route-level metrics. Perfect modelling of optionality is impossible without mark-remark data for all options. In order to evaluate the integrity of route and qualification metrics, further work is recommended on the sensitivity of the underlying models.

Appendix B – Multi-level model

The probability that a candidate is awarded the correct grade can also be derived by using a multi-level model to fit consistency of marking. The parameters from this model can then be used to simulate the mark-remark difference at component level for a set of randomly generated candidates. The algorithm for generating pseudo-candidates is discussed in more detail elsewhere (Schumann, E., 2009)[17] and so will only be briefly covered here.

The multi-level model relates the mark-remark difference for each item to the final mark awarded to the candidate for the item and the maximum mark of the item within each component within each subject. These variables were chosen on the basis that they were both likely to influence the level of agreement between examiners.

All components from a single subject from all exam boards are included in a single model. The model has been constructed with three levels. Marking events (*i*) are nested within questions (*j*) which are in turn nested within components (*k*). The model is given by the following equation

$$diff_{ijk} = \beta_{0jk} + \beta_1 final\ mark_{ijk} + \beta_2 maximum\ mark_{ijk} + e_{ijk}. \quad \text{(A1)}$$

For each question within a component, the final mark awarded is randomly generated for 5000 candidates. For each question the distribution of marks is roughly uniform and the correlation between questions is approximately 0.4 (Schumann, E., 2009). Equation A1 is then used to simulate the randomly generated mark-remark difference for each candidate and this simulation is replicated 25 times, giving the equivalent of 125,000 candidates. Each candidate has a final mark for every question on the component and a corresponding mark-remark difference. From this, the final mark awarded and mark-remark difference are calculated at component level for each candidate. For the final mark awarded to each candidate the mean difference and standard deviation are summarised from each of the 25 replications. Probabilities that a candidate has been awarded the correct grade are determined by their positions relative to the nearest grade boundary. Finally, the output statistic is the weighted mean of the probability that a candidate has been awarded the correct grade, where the weights are the number of pseudo-candidates at each mark.

[17] E. Schumann. (2009). Generating Correlated Uniform Variates. Retrieved from http://comisef.wikidot.com/tutorial:correlateduniformvariates

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/18/6449/2**