View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Digital Education Resource Archive

CAMBRIDGE ASSESSMENT

Estimation of inter-rater reliability

January 2013

Note: This report is best printed in colour so that the graphs are clear.

Vikas Dhawan & Tom Bramley
ARD Research Division
Cambridge Assessment

Ofqual/13/5260

This report has been commissioned by the Office of Qualifications and Examinations Regulation.

University of Cambridge Local Examinations Syndicate

Acknowledgements

We would like to thank our colleague Beth Black at OCR for her advice, and OCR for allowing access to their data.

Table of contents

Executive summary

The main aim of this research was to investigate estimates of inter-rater reliability for assessments where inconsistency in marking between markers might be thought to be the major source of unreliability.

Marker agreement was investigated by comparing awarded marks with the pre-determined definitive marks on seed scripts. Four pairs of on-screen marked units/components were selected from OCR's June 2011 session, two each from GCE and GCSE. Each pair had two units/components – one having long, essay-type questions (referred to here as 'Long components') and the other having objective-type questions (referred to here as 'Short components'). Apart from this difference, both units/components in a pair were chosen to be as similar as possible. The extent of difference in marker agreement between the Long and the Short components was used to investigate the effect of marker unreliability on examinees' scores. The two constituents of each pair were also compared in terms of internal reliability statistics like Cronbach's Alpha and SEM. It was assumed that if these indices appeared worse for the Long components, it would be an indication of the effect of marker inconsistency.

The main findings of the research can be summarised as:

- The spread of marker differences from the definitive mark was larger in the Long components than their corresponding Short components.

- The Short component in each pair had a lower standard error of measurement than its corresponding Long component, and in all but one of the pairs the Short component had a higher Cronbach's Alpha.

- In general the markers were on average neither too severe nor too lenient, though some slight variations were observed among the pairs.

- The marker differences were spread wider in the components which had a higher paper total.

- On average, marks awarded in the Short components were found to be closer to the definitive marks when investigated in a more fine-grained manner according to each seed script and each marker. More variation was observed in the Long components, some of which could be attributed to instances where most of the markers did not 'agree' with the definitive mark.

- Average marker differences were found to be within the tolerance level (defined here as a *range* of acceptable marker differences) in the Short components but appeared outside the tolerance levels for a greater proportion of markers in the Long components.

- The lower the maximum numeric mark of an item, the higher the level of marker agreement was found to be.

- A relatively crude method of analysing classification consistency suggested that all examinees in the Short components would be more likely to get the same grade if their work was marked by a different marker than if they sat a different (parallel) test, but this was less clearly the case for the Long components.

While the data used in this study came from the exam board OCR, the findings are likely to apply equally to equivalent types of examination from other awarding bodies.

1. Introduction

This report is based on the work commissioned by Ofqual under its reliability programme which is aimed at continuing its research into assessment reliability of results from national tests, public examinations and other qualifications that Ofqual regulates. The aim of this study was to investigate estimates of inter-rater reliability for assessments where inconsistency in marking between markers represents the major source of unreliability. The data used for this report was made available by the awarding body OCR from live high-stakes GCSE and A level examinations taken in England during the June 2011 session.

Ofqual has recently completed a two-year research programme that explored a range of issues associated with assessment reliability and produced a series of research reports. Cambridge Assessment contributed to the programme with its work on estimating reliability of qualifications (Bramley & Dhawan, 2010). One of the strands of that research was investigating marker-related variability in examination outcomes in GCSE and A level qualifications. The current study can be viewed as an extension to that strand.

One limitation of the previous report was that the analysis was restricted to the kinds of units and components[1] that had been marked on screen in the June 2009 examination session. These were mainly components consisting of short-answer questions where one might expect a higher level of marking accuracy. Papers comprising mainly long essay questions could not be used in the study. Since the previous report, however, more and more components have moved to being marked on screen, with the result that more on-screen marked long-answer and essay-type questions were available in the June 2011 session. For the current study, we have used components having more extended-response questions to investigate estimates of inter-rater reliability for assessments where inconsistency in marking between markers could represent the major source of unreliability.

Of course, which source is the major source of unreliability in any given set of scores depends on what kind of replication of the assessment is being envisaged. If a different (parallel) test was administered to the same group of candidates, error sources would include test items, markers, and potentially occasions as well. In a re-marking or multiple-marking study the only source of error is inconsistency among markers. The focus of this study was on comparing matched pairs of exam components where the members of each pair differed in type of question (long answer vs. short answer). The assumption was that differences between the members of each pair in indicators of reliability could be attributed to differences in reliability of marking, regardless of whether the indicator included various sources of error (Cronbach's Alpha) or a single source (multiple marking of seed scripts).

Different terminology has been used in the examination literature to conceptualise marking inconsistency between markers. Bramley (2007) suggested the use of the terms 'agreement' for questions that require longer or essay-type responses, 'accuracy' for multiple-choice or objective questions and 'reliability' for situations where marker inconsistency is conceived as a ratio of true variance to total (true + error) variance using either classical test theory (CTT) or item response theory (IRT). The focus of the current report was long/essay-type questions and therefore the term 'agreement' has been used to quantify the level of inconsistency in marking between markers, unless otherwise specified.

The investigation of marker inconsistency in this study was done by estimating the extent of agreement between marks given by markers with the marks decided by the Principal Examiner (PE) and senior examining panel to be the correct mark (i.e. the 'definitive' or the 'gold standard'

[1] 'Component' has been used in this report as a generic term for either a unit of a unitised assessment, or a component of a linear assessment, or a component of a unit of a unitised assessment. These distinctions are not of much relevance for this analysis.

mark) on 'seed scripts'. The seed scripts are the complete work of an examinee on the component for which the definitive mark on every question has been agreed by a panel of experts. A comparison of the marker's mark with the definitive mark indicates whether a marker is applying the mark scheme correctly. Inter-marker agreement was investigated using data from the monitoring of on-screen marking via seed scripts.

Seed scripts are inserted into the marking allocation of each marker at a given rate (e.g. 1 in 20). The great advantage of using seed scripts for investigating marker reliability is that each seed script is independently marked by all the markers on the panel – a 'blind multiple-marking' scenario which is very expensive to create in a separate research exercise. For a detailed description of the theory of marker-related variability and use of seed scripts for marker monitoring please refer to section 2 of Bramley & Dhawan (ibid.).

2. Selection of components

The data used for this study was made available by the exam board OCR from the June 2011 session. Eight components were selected for this research, which had all been marked on-screen[2]. The question papers and mark schemes from past sessions of various qualifications can be downloaded from OCR's website[3].

The screening of components to decide whether they were likely to be in the category of those where inconsistency in marking between markers represents the major source of unreliability was done on the basis of the maximum mark available for the questions in the component. The assumption was that higher-tariff questions were more likely to be essay questions, or questions with relatively complex mark schemes. The components selected under this category (referred to as 'Long' components in this report) had at least one item (i.e. sub-part) which was worth eight marks or more.

Marker agreement in the Long components was compared with marker agreement in the components where inconsistency in marking was deemed to be comparatively lower. Under this category (referred to as 'Short' components in this report), only those components were selected in which each item was worth less than eight marks.

Four pairs of components were used in the analysis, two each from GCE and GCSE qualifications. In each pair one member contained questions that were likely to be marked with a high degree of accuracy/agreement (i.e. Short components) whereas the other member of the pair contained at least some questions where we might expect markers to differ (i.e. Long components). The extent of difference in marker agreement between the two categories was used to indicate the effect of marker unreliability on examinees' scores.

For each pair, only those components were selected which had the same paper total (maximum numeric mark). The target was to obtain pairs that were as similar as possible in terms of:

- number of markers
- number of seed scripts
- grade bandwidth
- raw score distribution.

The purpose of this close matching of pairs was to try to ensure that as far as possible any differences in statistics relating to marker agreement should be attributable to differences in reliability of marking of the types of question involved. It is therefore the comparison between members of each pair that are of most interest in this report.

[2] For on-screen marking OCR uses the Scoris™ system in partnership with RM plc.
[3] For instance, see http://www.ocr.org.uk/qualifications/type/gce/blt/accounting/documents/ (accessed on 5th December, 2011).

An additional criterion applied to select components was that a component should have at least 10 different seed scripts and at least five different markers. Assessment material like question papers and mark schemes were also consulted so as to include components which had more essay-type questions from the available Long components.

The four pairs, selected in consultation with Ofqual, are given in Table 2.1. The table shows a Long component with a matching Short component in each pair. For this report, the components have been given a label according to their pair number and type (for instance: 1L=Pair 1, Long component and 4S=Pair 4, Short component).

Table 2.1 Selected pairs, June 2011 session

Pair Number	Qualification	Type	Component Label
1	GCE	Long	1L
1	GCE	Short	1S
2	GCE	Long	2L
2	GCE	Short	2S
3	GCSE Unit	Long	3L
3	GCSE Unit	Short	3S
4	GCSE Unit	Long	4L
4	GCSE Unit	Short	4S

3. Internal consistency estimates

This section gives the results obtained from comparing each member of the pair in terms of internal reliability statistics such as Cronbach's Alpha, Standard Error of Measurement (SEM) and ratio of grade bandwidth to SEM using data from all examinees. Grade bandwidth here refers to the difference between the A to B or C to D boundary. It was assumed that if these indices appeared worse for the Long component in the pair, it would be an indication of the effect of marker inconsistency. Table 3.1 gives the summary statistics of the components.

The table gives the value of Cronbach's Alpha, which is the most widely reported statistic of internal consistency of a test. Along with it, the SEM given in the table gives an indication of the precision of measurement of the tests. The lower the SEM (and the higher the Cronbach's Alpha), the more reliable the test instrument is generally accepted to be. The table also gives another measure, Bandwidth:SEM ratio, which was introduced in Bramley & Dhawan (ibid.). The authors argued that the use of this ratio could allow for more meaningful comparisons between components because Cronbach's Alpha and SEM cannot be properly interpreted without taking into account the maximum mark of the test or the number of items. In this study, although the maximum mark of both the components in a pair was equal, the Bandwidth:SEM ratio was relevant because SEM is given in raw marks whereas the final outcome is reported in grades. The use of this ratio allows comparison in terms of grades. The higher the ratio, the more repeatable the grade outcomes are likely to be.

This concept can also be explained as the probability of a person with a true score in the middle of the grade band getting a grade outside the band. The probability values for the selected components are given in Table 3.1 in the column *Prob. Outside*. Ideally, we would want to have this value as low as possible.

For a more detailed explanation of these concepts, please refer to section 1 in Bramley & Dhawan (ibid.).

Table 3.1: Summary statistics of components, June 2011

Pair Num	Type	Comp. Label	Qualification	Paper total	# Items	Entry size	Grade Bandwidth	Grade Range	Mean	SD	Cronbach's Alpha	SEM	Bandwidth: SEM	Prob. Outside
1	Long	1L	GCE	90	12	7246	7	AB	40.47	12.06	0.75	6.05	1.16	0.56
1	Short	1S	GCE	90	40	5188	7	AB	52.16	16.32	0.91	4.92	1.42	0.48
2	Long	2L	GCE	60	13	12466	4	AB	37.23	9.45	0.74	4.77	0.84	0.68
2	Short	2S	GCE	60	30	8827	5	AB	38.41	10.87	0.88	3.75	1.33	0.51
3	Long	3L	GCSE Unit	60	19	3060	6	AB	36.67	9.33	0.84	3.73	1.61	0.42
3	Short	3S	GCSE Unit	60	34	3620	6	CD	30.92	6.63	0.76	3.28	1.83	0.36
4	Long	4L	GCSE Unit	60	19	6847	7	AB	33.51	10.46	0.80	4.72	1.48	0.46
4	Short	4S	GCSE Unit	60	43	11428	8	AB	31.43	10.23	0.88	3.49	2.29	0.25

Items = Number of items (i.e. sub-parts) in the question paper
Grade bandwidth = marks in the A-B or C-D range
Mean/SD = Mean and standard deviation of the marks obtained by all examinees
SEM = Standard Error Measurement
Bandwidth:SEM = ratio of grade bandwidth and SEM
Prob. Outside = Probability of a person with a true score in the middle of the grade band getting a grade outside the band.

A comparison of Cronbach's Alpha within each pair is also shown in Figure 1. In the figure, blue dots represent the Short components whereas the Long components are shown by red dots[4]. The figure also shows the paper total, which was equal for both the components in each pair. Figure 1 shows that in all the pairs except Pair 3, the Long component had a lower value of Cronbach's Alpha than its corresponding Short component. Overall, the value of Cronbach's Alpha was fairly high for all the components (approximately 0.75 or above). The absolute difference in Cronbach's Alpha between the Long and the Short component was larger in pairs 1 and 2 (GCE) than in pairs 3 and 4 (GCSE).

A comparison of SEM in the component-pairs is shown in Figure 2. Across all the pairs, the Long component had a higher SEM than the Short component. The difference in SEM between the corresponding Long and Short components was similar across all the pairs except Pair 3 in which the SEMs of the components were closer to each other. The components in Pair 1 had the highest SEM values, which was probably because they had the highest paper total.

As mentioned earlier, comparison of components based on SEM might not give an accurate picture of their relative precision. A comparison of the component-pairs using SEM given in Figure 2 and Bandwidth:SEM ratio given in Figure 3 was used to gain further understanding. In Pair 1, the difference in SEM between the Long and the Short components appears comparatively high which might lead to the interpretation that there is a wide gap in the precision of the two components. Figure 3, on the other hand, shows that the precision estimates of the two components in Pair 1 appear closer to each other using the Bandwidth:SEM ratio. This suggests that the two components might be having similar levels of measurement precision in terms of the grade scale. The highest difference in Figure 3 amongst all the pairs was observed in the Long and Short components of Pair 4 which indicated a higher level of precision for the Short component in comparison to its corresponding Long component.

From these internal consistency reliability statistics it appeared that the Short component in each pair had a higher precision of measurement than its corresponding Long component. All the components selected in this study were *single* units/components of larger assessments. Overall (composite) reliability of the whole assessment is likely to be higher as shown in section 1 in Bramley & Dhawan (ibid.).

[4] This colour pattern of blue for Short and red for Long components is followed throughout the report.

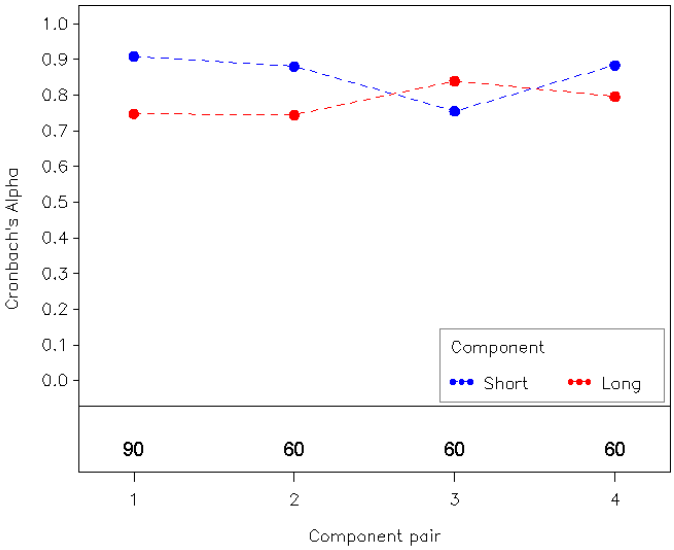


Figure 1: Cronbach's Alpha

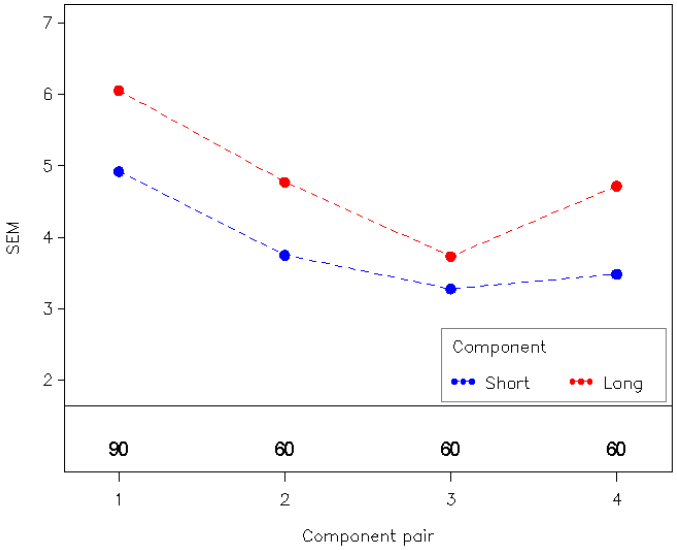


Figure 2: SEM

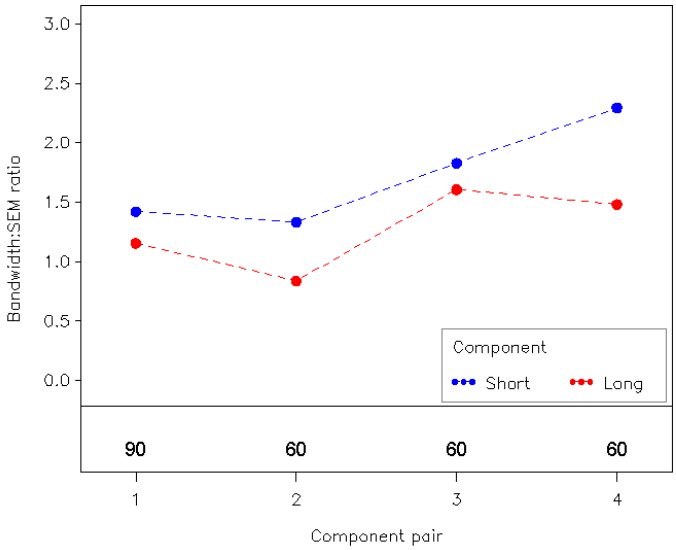


Figure 3: Bandwidth:SEM ratio

4. Marker agreement

This section reports the extent of differences between the marks awarded by markers and the definitive marks. As mentioned earlier, definitive marks are the marks decided by the PE and senior examining panel on seed scripts. The difference between awarded mark and definitive mark was used in this study as a measure of marker agreement.

An overview of the seed scripts used in the study is given below in Table 4.1.

Table 4.1: Summary statistics of definitive marks of seed scripts, June 2011

Pair Num	Type	Qual.	Comp. Label	Paper Total	# Seed Scripts	Mean	SD	Median	Max	Min
1	Long	GCE	1L	90	10	45.13	8.04	44	60	35
1	Short	GCE	1S	90	15	68.47	8.57	67	82	43
2	Long	GCE	2L	60	15	36.04	11.00	38	55	15
2	Short	GCE	2S	60	23	46.21	6.91	49	55	31
3	Long	GCSE Unit	3L	60	15	37.46	6.78	38	50	23
3	Short	GCSE Unit	3S	60	20	32.75	3.17	32	39	28
4	Long	GCSE Unit	4L	60	15	35.00	8.15	36	49	20
4	Short	GCSE Unit	4S	60	20	32.06	7.77	31	53	17

Seed Scripts= Number of seed scripts in a component

Table 4.1 shows that the mean (and median) of definitive marks was higher for the Short component in pairs 1 and 2 and higher for the Long components in pairs 3 and 4. A comparison of Table 4.1 with Table 3.1 (which gives summary statistics of the marks obtained by *all* examinees) shows that the mean marks of all examinees were similar to the mean (definitive) marks on the seed scripts except in components 1L, 1S and 2S. The mean marks in these three components were comparatively higher for seed scripts. A comparison of the standard deviation in the two tables shows that definitive marks had a narrower spread in all components except component 2L, where the reverse was true.

We would not necessarily expect the seed scripts to be representative of the whole distribution in a statistical sense, because very low scoring scripts with large numbers of omitted answers are unlikely to be chosen as seed scripts.

4.1 Overview of marker differences

Table 4.2 gives the summary of actual (i.e. signed) differences between awarded mark and definitive mark in seed scripts.

The table shows the number of seed scripts, markers and items in each component. The number of marking events (#MEs) gives the number of instances where a seed script was marked by a marker. The mean, standard deviation and median of the actual differences are also given in the table. The inter-quartile range (IQR) and the 5^{th} and the 95^{th} percentile give an idea of the spread of the differences. The table gives the correlation between awarded mark and definitive mark across all marking events.

The mean and median of the differences were close to zero for all the components which suggests that the markers were neither too lenient nor too severe. A positive value indicates leniency, thus the marking of the seed scripts in the Long component of Pair 1 and both the components of Pair 4 was lenient on average compared to the definitive mark. The Long component of Pair 4 had the largest value of the mean difference and was also the only component with a non-zero (positive) median value. In all the pairs the Long components had a

larger standard deviation (and inter quartile range) than their corresponding Short components, indicating greater fluctuation around the mean (and median). The largest values of these measures were observed in the Long component of Pair 1, which was not an unexpected finding given the fact that the components in this pair had the longest mark range.

The correlation between definitive marks and awarded marks was fairly high in all the components. In all the pairs, the correlation was higher for the Short component than the Long component. In Pair 2, this difference was found to be very small. Correlations can be a misleading indicator of agreement and have been given here only for the purposes of comparison with other work on marker reliability. The distribution of actual differences (given later in Table 4.3) is a more informative representation of marker agreement.

In addition to the summary of distribution of differences between definitive mark and awarded mark given in Table 4.2, the last column of the table gives the median of inter-correlations between marks awarded by markers on seed scripts. This statistic[5] gives an estimate of the consistency of marks *among* the markers as opposed to the comparison with the definitive marks. The median of the correlations was high (and similar to the overall correlations between awarded and definitive mark) for all the components except 1L. In this component, the markers did not seem to agree to a great extent with each other in their assessment of candidate performance. It should be emphasised that the number of data points for calculation of these inter-marker correlations was very small, being limited by the maximum number of seed scripts available in a component.

Table 4.3 shows the percentage of marker differences in different categories with differences ranging from less than -7 to greater than +7. The table also shows the proportion of these differences which were within the grade bandwidth of the component.

The table shows that the largest percentage of marker differences in all the components was within the -1 to +1 range. It is striking that in all pairs the percentage of differences in the -1 to +1 range was more than twice as high for the Short component than for the Long component. The largest proportion of differences within the grade bandwidth was observed in the component 3S whereas component 1L had the lowest proportion.

[5] Where a marker had marked a seed script more than once, only the first instance of marking was included in the calculation of correlations.

Table 4.2: Summary of distribution of differences between definitive mark and awarded mark for seed scripts

Pair Num	Type	Qual.	Comp. Label	# scripts	# markers	# items	Paper Total	# MEs	Mean	SD	Median	IQR	P5	P95	Corr.	Median (inter-marker corr.)
1	Long	GCE	1L	10	33	12	90	372	0.29	7.43	0	9	-12	13	0.89	0.55
1	Short	GCE	1S	15	24	40	90	246	-0.09	1.97	0	2	-3	3	0.97	0.97
2	Long	GCE	2L	15	41	13	60	636	-0.14	3.68	0	5	-6	6	0.97	0.94
2	Short	GCE	2S	23	28	30	60	452	-0.02	1.17	0	2	-2	2	0.98	0.98
3	Long	GCSE Unit	3L	15	11	19	60	145	-0.78	3.51	0	4	-6	5	0.89	0.93
3	Short	GCSE Unit	3S	20	11	34	60	184	-0.01	0.75	0	0	-1	1	0.99	0.96
4	Long	GCSE Unit	4L	15	24	19	60	350	0.59	3.84	1	5	-6	7	0.93	0.90
4	Short	GCSE Unit	4S	20	30	43	60	593	0.24	1.13	0	1	-1	2	0.98	0.98

Key: # items= number of part-questions on the exam paper. # MEs= number of 'marking events' where a seed script was marked by a marker. Includes repeated markings of the same seed script by the same marker. IQR= Inter-quartile range. P5/P95= 5^{th}/95^{th} percentile. Corr.= Pearson correlation between awarded mark and definitive mark across all marking events. Median (inter-marker corr.)= median of *inter-marker* correlations of marks awarded by markers.

Table 4.3: Distribution of differences between definitive mark and awarded mark for seed scripts

Pair Num	Type	Qual.	Comp. Label	Paper Total	# MEs	<-7	-7 to -5	-4 to -2	-1 to +1	+2 to +4	+5 to +7	>+7	Grade Bandwidth	% within grade bandwidth
1	Long	GCE	1L	90	372	14.5	8.1	13.2	23.4	12.9	12.9	15.1	7	41.7
1	Short	GCE	1S	90	246		2.4	19.5	59.8	17.5	0.8		7	92.7
2	Long	GCE	2L	60	636	2.2	8.6	23.0	36.8	18.6	9.0	1.9	4	44.5
2	Short	GCE	2S	60	452			8.2	83.2	8.6			5	95.6
3	Long	GCSE Unit	3L	60	145	0.7	15.9	25.5	36.6	15.2	4.1	2.1	6	57.9
3	Short	GCSE Unit	3S	60	184			2.7	95.7	1.6			6	99.5
4	Long	GCSE Unit	4L	60	350	2.0	6.6	18.9	34.0	25.7	9.1	3.7	7	67.1
4	Short	GCSE Unit	4S	60	593			4.7	84.0	11.1	0.2		8	99.6

The marker differences of each component-pair are also shown as a box plot in Figure 4. In the figure, the horizontal line inside the boxes represents the median of the differences. The length of the box represents the interquartile range (from 25th to the 75th percentile). The T-lines extended from each box show the 5th to 95th percentile range. The horizontal line at 0 represents the line of no difference i.e. the point where awarded mark is equal to the definitive mark of a seed script.

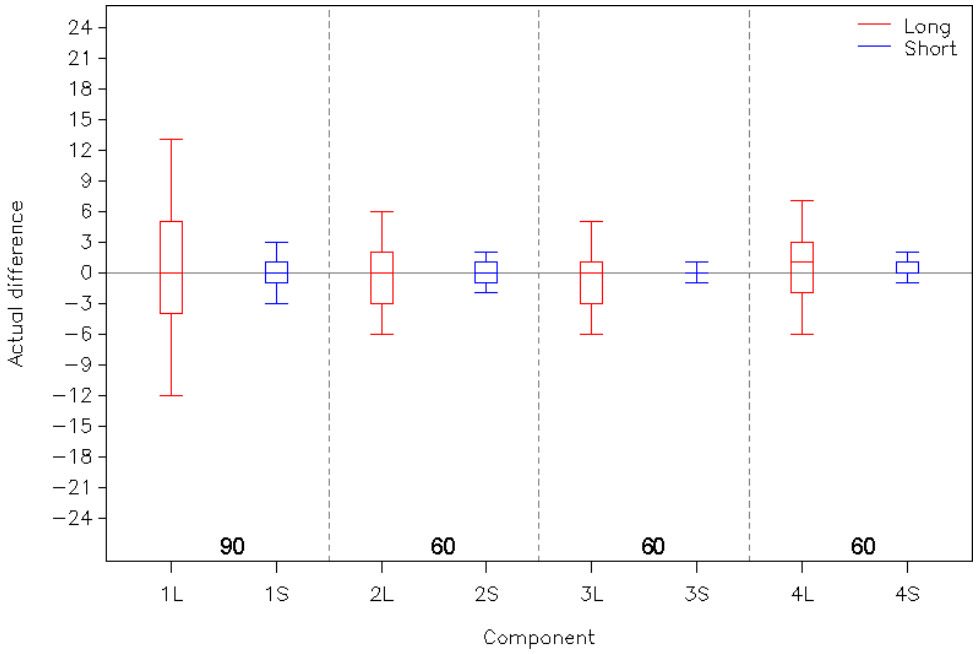


Figure 4: Box plot of the distribution of differences between awarded and definitive mark across all markers and seed scripts in each pair.

Figure 4 shows that in all the pairs the spread of marker differences was larger in the Long component than its corresponding Short component. The Long component in Pair 1 had the largest spread of marker differences, as was noted earlier. Also, Pair 1 had the largest difference in the spread of marker differences between the corresponding Long and Short components amongst all the pairs. This difference between the Long and the Short components appeared to be more or less the same across the rest of the pairs.

In both the components of Pair 4, a greater proportion of marker differences were above the line of no difference. This suggested more lenient marking in these components. Also, component 3L had a greater proportion of marker differences below the line of no difference, which suggested that markers were slightly more severe in this component. Overall the spread of differences appeared similar above and below the line of no difference, which indicated that, as mentioned earlier, markers were neither too severe nor too lenient.

The summary of marker differences given above suggested that overall the markers were neither too severe nor too lenient. Differences were spread wider in components with higher paper total (i.e. in Pair 1). In all the pairs the spread of marker differences was larger in the Long component than its corresponding Short component.

4.2 Comparison of marker differences for each pair

This section gives a more detailed view of how the marker differences varied between the Long and the Short components in each pair. The information given here is effectively the same as given in Figure 4, though in more depth.

Figure 5 (a to d) shows histograms of differences in the two components for each pair.
The figures also give some summary statistics of the marker differences. N here represents the number of seed script marking events (also given in Table 4.2).

The graphs in Figure 5 show that the highest concentration of differences for all the components was around the ‘0’ on the x-axis, which represents a point of complete agreement between awarded and definitive marks on seed scripts. This concentration was more pronounced for the Short components in all the pairs. The Long components had more flattened bars and a wider spread of differences than their corresponding Short components.

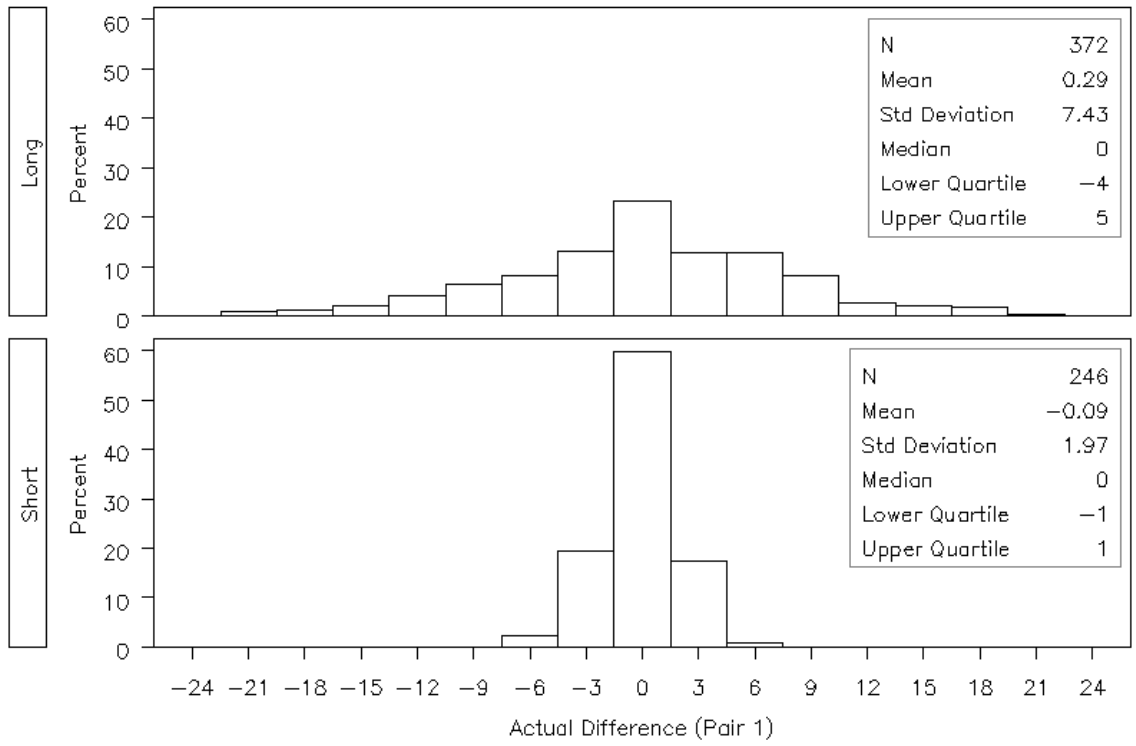


Figure 5a: Marker differences, Pair 1

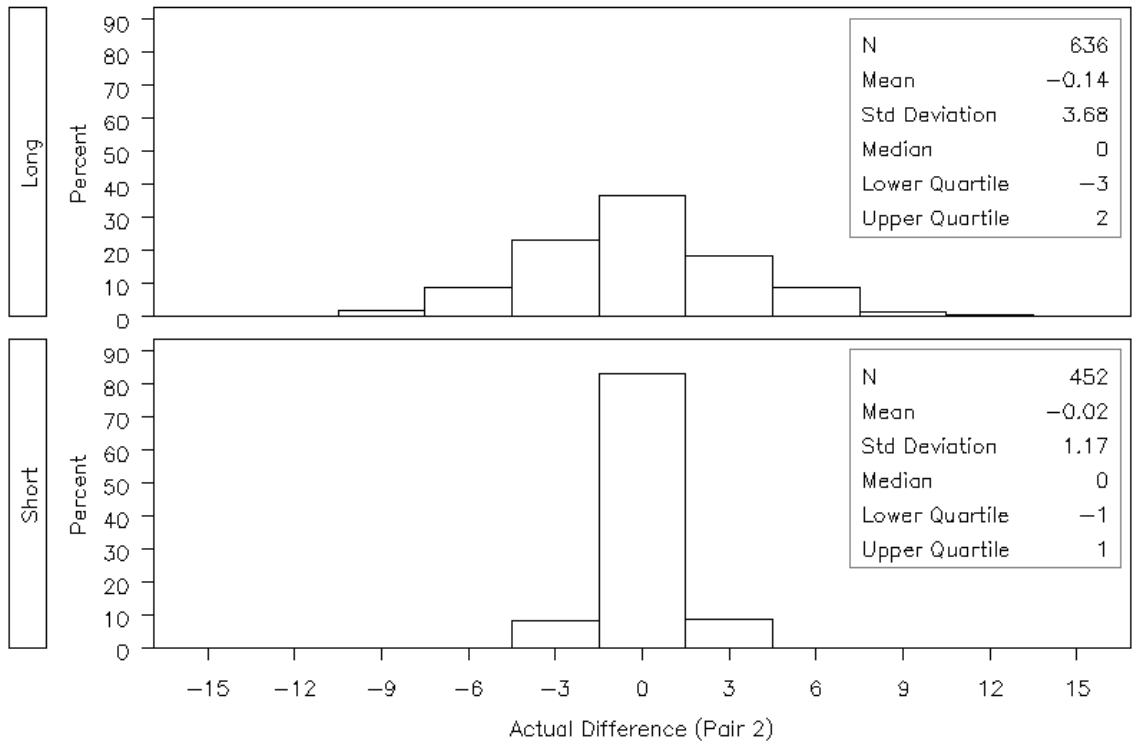


Figure 5b: Marker differences, Pair 2

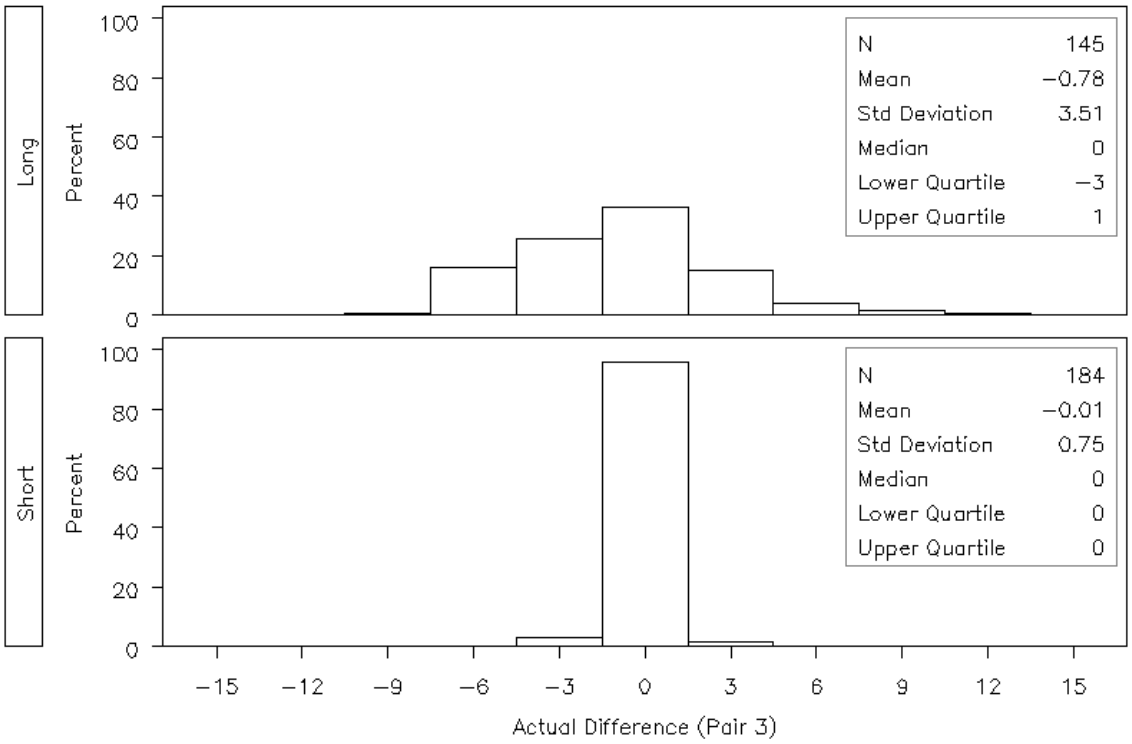


Figure 5c: Marker differences, Pair 3

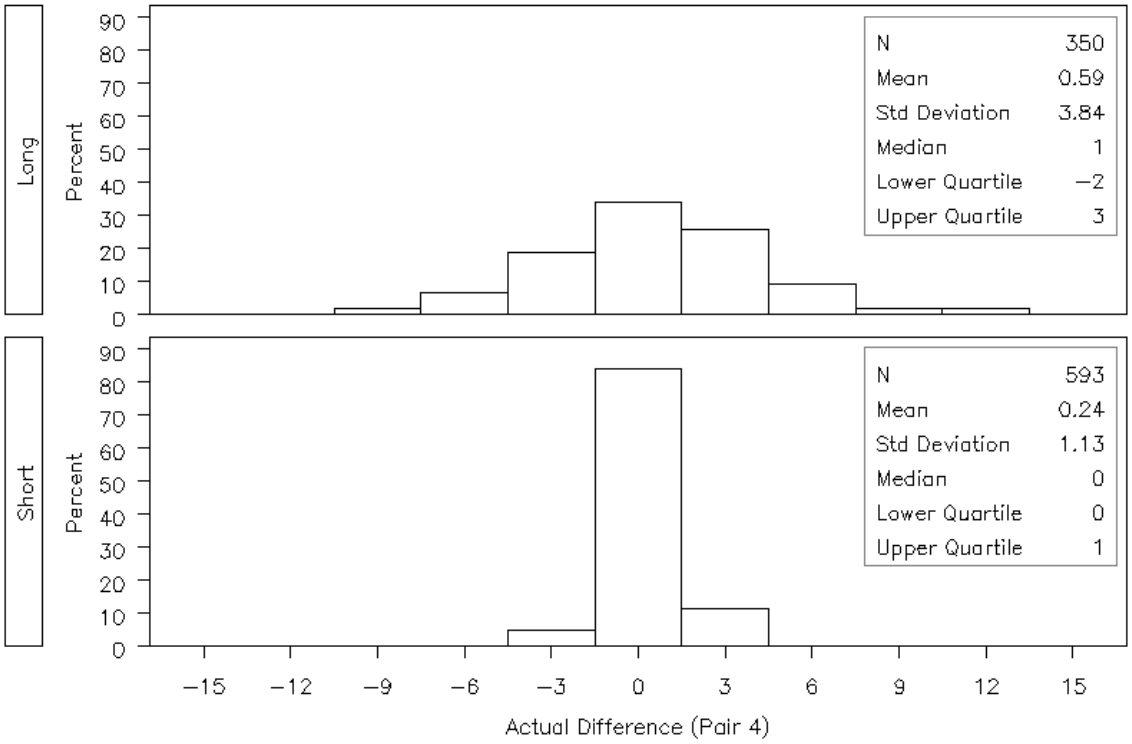


Figure 5d: Marker differences, Pair 4

The graphs in Figures 4 and 5 treat all differences the same regardless of where they occurred on the mark scale. In order to see whether there was a tendency for more or less agreement at different parts of the mark scale, the standard deviation of actual differences was plotted against the definitive marks of seed scripts for all the eight components (shown in Figure 6).

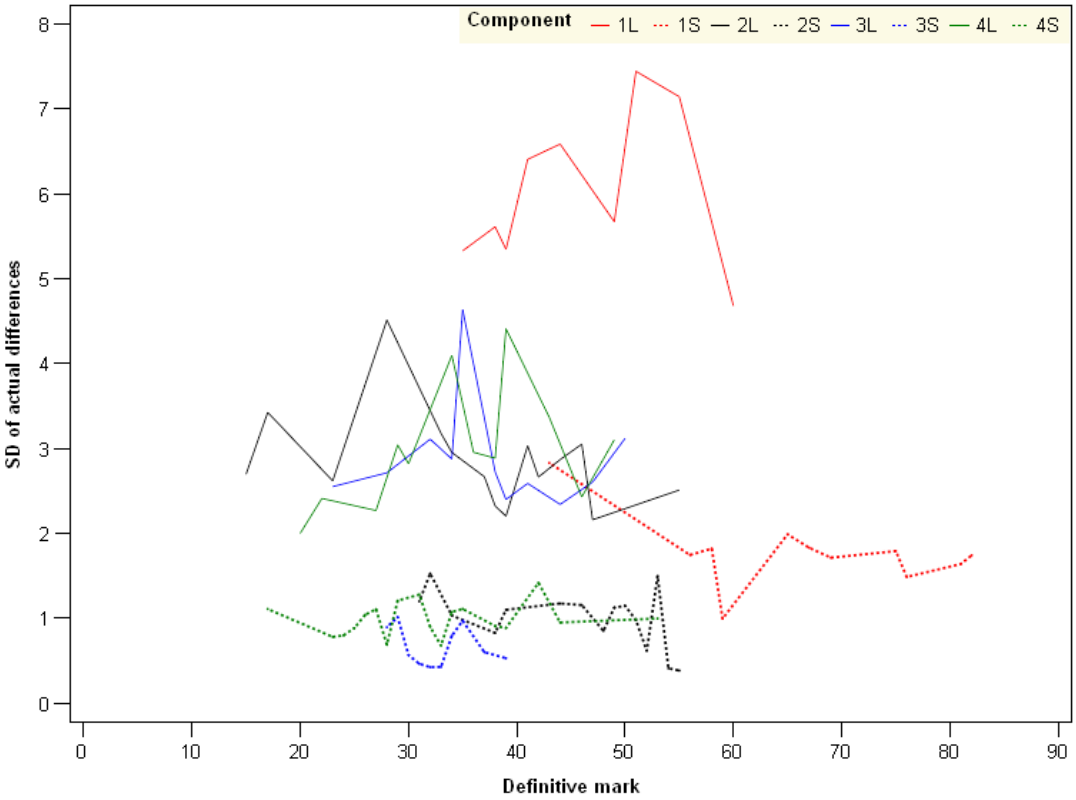


Figure 6: Spread of actual differences according to definitive mark, all components.

Figure 6 shows that, as noted earlier, the spread of marks was larger for the Long components (shown by solid lines). Overall, no consistent trend was observed between the standard deviation of the differences and the definitive mark. There were slight variations from this trend in component 4L (solid green line) where the spread of differences appeared to increase and in component 2S (dotted black line) where the spread appeared to decrease, on an average, with the increase in the definitive mark. However, in general it did not appear that the spread of differences increased or decreased consistently with the increase in the definitive marks of scripts.

Figure 7 shows marker differences for all the pairs in a more fine grained manner - by each seed script. In this figure, the differences between the awarded and the definitive mark for each marker on each seed script are shown for all the components. The red or blue dots show the differences according to each script marking instance in the Long or the Short components respectively. The black dots connected by a line show average (mean) differences on each seed script.

The x-axis in Figure 7 shows the sequence number of seed scripts in a component. The scripts have been ordered by their total definitive mark, from low to high. The line at 0 on the y-axis shows the line of no difference (complete agreement at the whole script level between the awarded mark and the definitive mark of seed scripts). Differences above this line indicate lenient marking whereas those below the line indicate severe marking.

The lines representing average differences appear to more or less overlap the line of no difference in the Short component for all the pairs. This suggests that the marks awarded for all the scripts in Short components were, on average, very close to their definitive marks. There was more variation in the Long components where particular scripts like #8 in Pair 1 and #6 in Pair 4 showed a large average disagreement with the definitive mark. A large average difference on either side of zero could arise if there was a lot of disagreement among the markers, but could also occur if most of the markers agreed with each other but disagreed with the definitive mark, which, in fact, was the case in these two scripts. Figure 7 (Pair 1, Long component) shows that in this component almost all the markers gave lower marks on script #8 than its definitive mark. Therefore most of the red dots on the graph are below the line of no difference. The mode of the marks given by markers was different from the definitive mark, which suggests that the definitive mark might not be the ‘correct’ mark on these scripts. Black et al. (2010) introduced the term ‘DIMI’ (definitive mark incongruent with modal mark items) to describe item marking instances where the majority of the markers did not agree with the definitive mark as the ‘correct’ mark. A higher proportion of DIMIs could be expected in components with essay-type questions than those having objective questions, which might result in a disagreement between the definitive mark and the modal mark on the whole script as well.

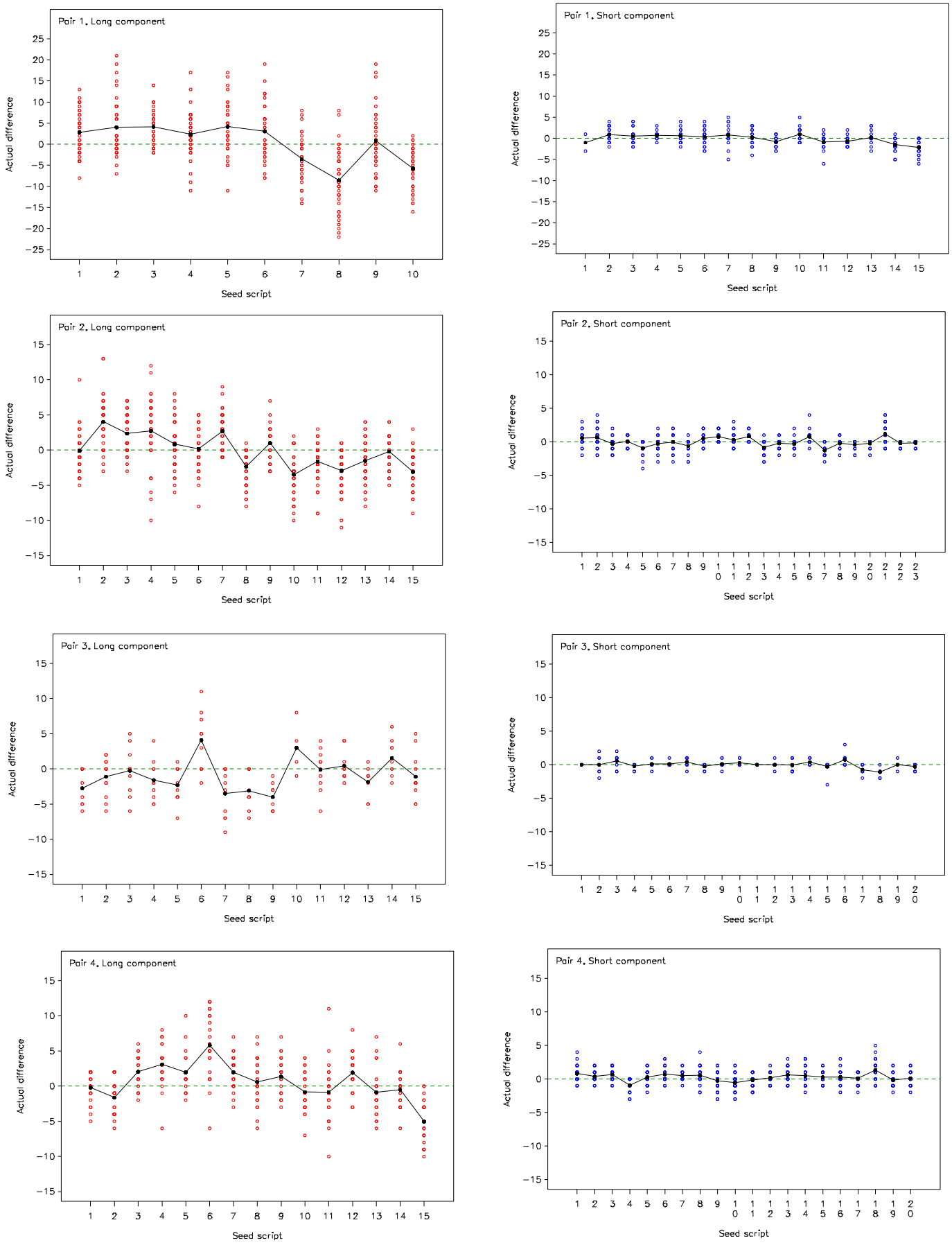


Figure 7: Actual difference (across markers) between awarded and definitive mark, displayed for each seed script. The black dots connected by a line give average (mean) differences.

The information in Figure 7 can be re-organised to show the distribution of differences from each marker (across all their seeding scripts). This is displayed in Figure 8 below. Graphs like these can help in monitoring of markers by exam boards. Note that the markers have been listed in no particular order.

As in Figure 7, the line at 0 on the y-axis shows the line of no difference (complete agreement at whole script level between the awarded mark and the definitive mark of seed scripts). Differences above this line indicate lenient marking whereas those below the line indicate severe marking.

Figure 8 shows that in the Short components almost all the markers, on average, were neither severe nor lenient across the seed scripts allocated to them. This is represented by the close overlap of the lines of average marker differences with the green horizontal lines of no difference. On the other hand more average variation was observed in the average differences of markers in the Long components in all the four pairs.

As mentioned earlier, these graphs can help to study the performance of each marker. For instance, the graph of Pair 1 (Long component) shows that the highest amount of deviation from the definitive marks was observed in markers #10, #12 and #26. The direction of the deviation shows that markers #10 and #26 were more lenient whereas marker #12 was more severe than the rest.

A comparison of the Long components between the four pairs indicates that the markers in Pair 4 awarded marks which were on average closest to the definitive marks, followed by the markers in Pair 2 and Pair 3. Marker #33 in Pair 2 stood apart with the highest amount of (negative) deviation in the pair. As shown in the graph, this particular marker had marked only one seed script; so this does not give a reliable indication of their severity. It is likely the marker was stopped from marking given the large deviation and therefore did not continue to mark their full allocation of scripts.

Overall these figures also show that, at individual marker level, marks awarded to seed scripts in the Short components were closer on average to the definitive marks than in the Long components.

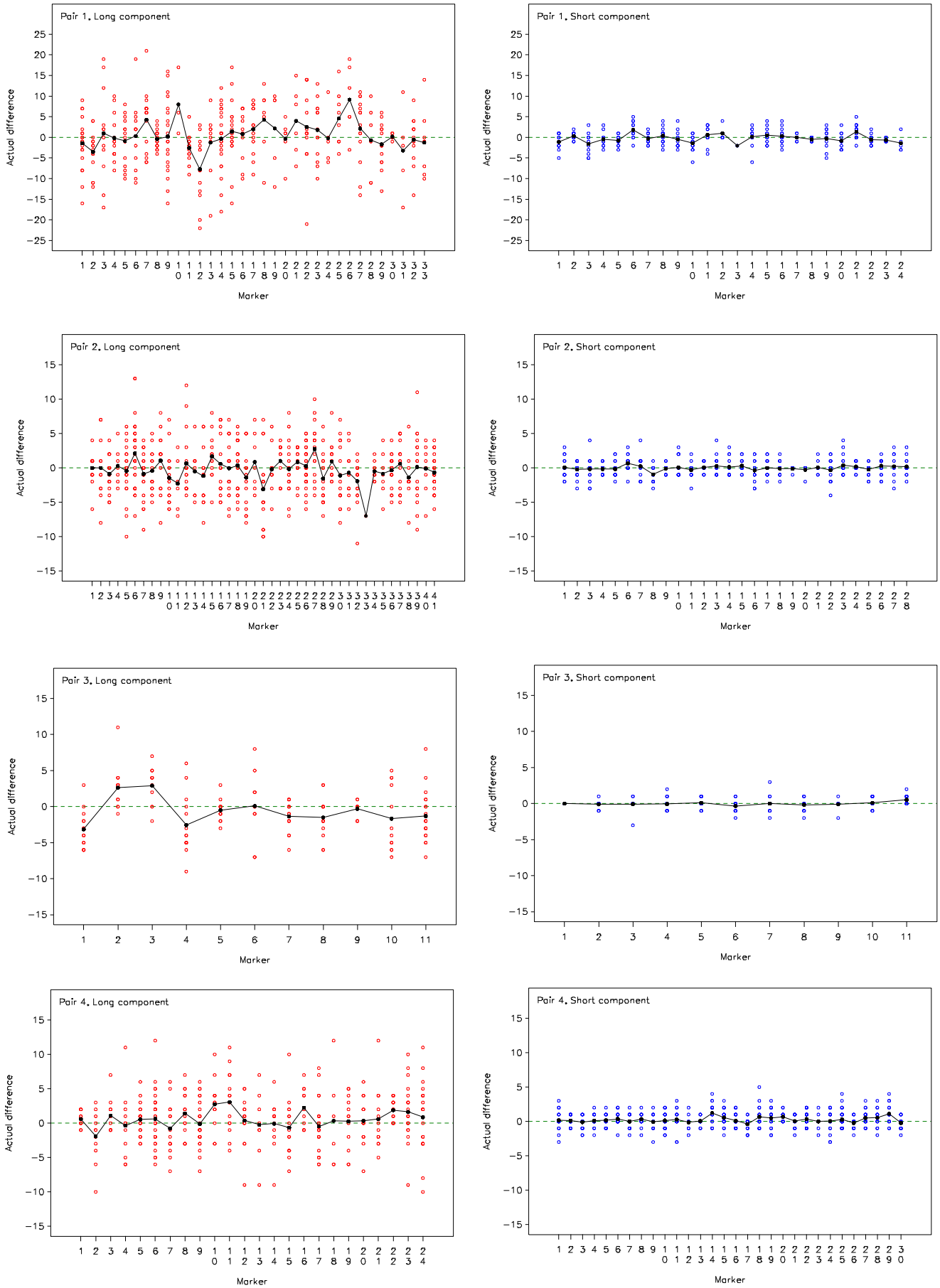


Figure 8: Actual difference (across seed scripts) between awarded and definitive mark, displayed for each marker. The black dots connected by a line give average (mean) differences.

4.3 Tolerance values

In the results given above, any variation from the definitive mark of a seed script was presented as a discrepancy in marking. However, comparing the extent of marker agreement in the Long and the Short components on this basis might be somewhat unfair. In the components which require more long answers and essay-type responses the markers might have to apply complex level-based mark schemes to interpret and judge candidate responses. In addition, the markers have to interpret mark schemes and decide the correct mark without having the advantage of participation in the extensive discussions which the PE and the senior examining panel might have had when deciding upon the definitive marks during the standardisation set-up meetings.

It would therefore seem more appropriate to compare the difference between awarded marks and definitive marks with a value representing the amount of 'tolerance' or acceptable deviation in awarded marks from definitive marks. OCR uses this concept[6] for monitoring of markers through seed scripts. Each question paper is allocated a tolerance value and if the sum of *absolute* (i.e. unsigned) differences across all the questions on a script[7] exceeds the tolerance value the marking instance is flagged up. Each marker is monitored using this process for each seed script marked.

The actual value of tolerance for each component is usually decided as a certain percentage of its paper total. Table 4.4 gives the tolerance values for the eight components used in this study.

Table 4.4: Tolerance values

Pair Num	Qualification	Type	Component Label	Paper Total	Tolerance value
1	GCE	Long	1L	90	10
1	GCE	Short	1S	90	5
2	GCE	Long	2L	60	4
2	GCE	Short	2S	60	3
3	GCSE Unit	Long	3L	60	4
3	GCSE Unit	Short	3S	60	3
4	GCSE Unit	Long	4L	60	4
4	GCSE Unit	Short	4S	60	3

The tolerance values given in Table 4.4 are shown in the graphs of all the pairs in Figure 9 (as two green horizontal lines), which gives the absolute differences according to each marker. The upper horizontal line represents the tolerance value for the Long component and the lower line represents tolerance for the Short component in each pair. The red and blue dots joined by lines represent average absolute differences across all the seed scripts marked by each marker in the Long and the Short component respectively. '0' on the y-axis represents exact agreement between the awarded and the definitive mark for every question on all seed scripts marked by the given marker.

Note that in Figure 9 the markers have been listed in no particular order and that the Long and the Short components in a pair might not necessarily have had the same number of markers. Also, the marker numbers have been used merely as identifiers for producing the graphs and therefore the differences should be compared as a trend only. For instance, in Pair 1 in Figure 9,

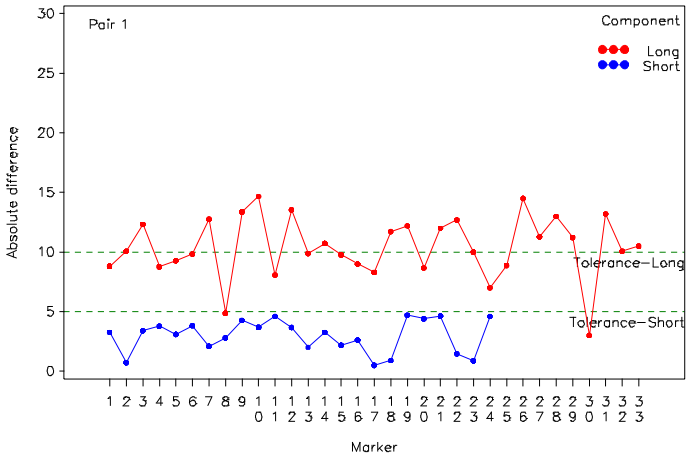
[6] OCR refers to this tolerance value as 'Scoris Variance' where Scoris is the software used for standardisation and marking.
[7] Referred to as 'Total Deviation' by OCR.

there is no particular interest in comparing marker #1 of the Long component with marker #1 of the Short component.

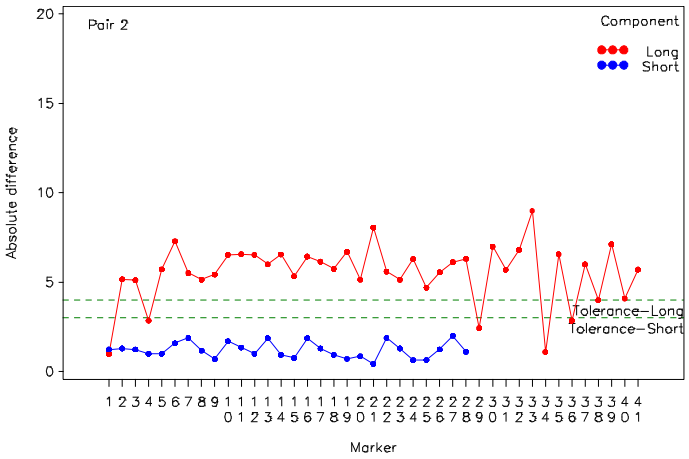
Figure 9 indicates that the average unsigned (or absolute) differences for all markers in all the Short components were within the tolerance levels. In the Long components, on the other hand, mixed results were found. A greater proportion of the average differences in these components appeared to be outside the tolerance levels. An exception to this was observed in the Long component of Pair 1 where average differences were either within or close to the tolerance level for a large number of markers.

This investigation of the use of tolerance values for marker monitoring indicated that, on average, marking was within tolerance for the Short components but outside tolerance for a higher proportion of markers in the Long components. However, this raises the question of the appropriateness of the tolerance values. As given in Table 4.4, in Pairs 2, 3 and 4, the difference between the tolerance values of the Long and the Short components was only one mark. This left a very narrow range of 'extra' tolerance available for the Long components as compared to the Short components having the same paper total. It could be argued that setting the tolerance value at a slightly higher percentage of the paper total in the Long components of these three pairs might have given more fair marker-monitoring results. In addition, giving extra weighting to other factors like complexity of the mark scheme and length of answers required might be of help as well. Having said that (without any intention to retrofit the solution or taking away the credit from the markers in the Long component of Pair 1!), setting tolerance values at too high a level is likely to be detrimental to its very purpose.

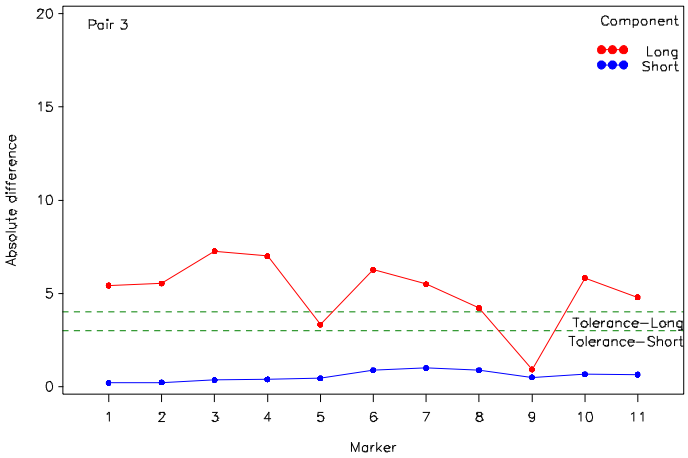
A rationale for setting tolerance values would be highly desirable. Black, Suto & Bramley (2011) present a review of the effect on marker agreement of certain features of questions, mark schemes and examinee responses. The application of some of their findings might help to set more realistic tolerance values at the item level. Other interesting recent work is that of Benton (2011), who approaches the problem of how to set optimum tolerances using a probabilistic model.



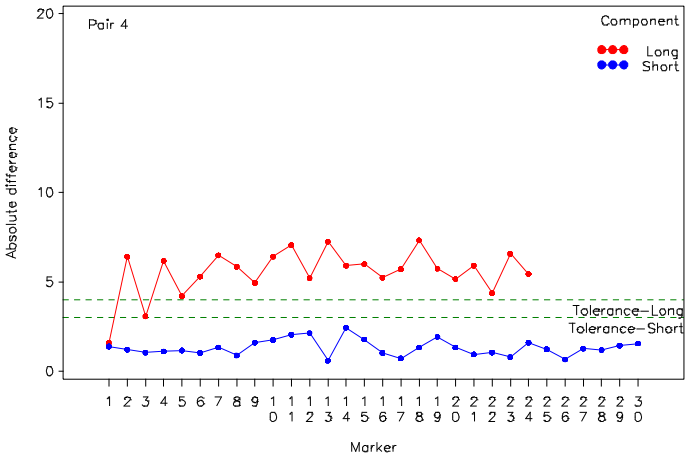
Pair 1



Pair 2



Pair 3



Pair 4

Figure 9: Average absolute difference (across seed scripts) between awarded and definitive mark, displayed for each marker. 'Tolerances' for the Long and the Short components also shown.

4.4 Alternative definitions of the 'definitive' mark

In the previous sections marker agreement was defined based on the difference between the awarded and the definitive mark on seed scripts. However, other definitions of 'definitive' marks are also possible, and are discussed below.

The mean awarded mark on the seed scripts is perhaps the most obvious alternative, with its connection with usual conceptions of 'true score'. Figure 7 shows how the awarded marks were distributed around this mean (the black lines) at the whole script level.

However, the mean awarded mark is arguably not appropriate as a definitive mark because it is usually not a whole number. The median can suffer from the same problem with an even number of observations. The mode avoids this problem, but in situations with a small number of markers and/or a large total mark for the paper, it is possible that there will be no mode (i.e. all the markers have a different total score for the seed script), or that the mode will relatively arbitrarily reflect a chance coincidence of the total marks given by a small number of markers.

If the concept of a 'correct' mark (see Bramley & Dhawan, 2010) for a script is useful, then this 'correct' mark is logically the sum of the 'correct' marks on each item. These correct item marks must also be whole numbers (except in the very rare cases of mark schemes that award half-marks, which was not the case for the components studied here). On the assumption that the mode of the awarded marks at the item level is most likely to be the 'correct' mark (which is certainly plausible in cases where careless errors or specific misunderstandings of the mark scheme by individual markers lead to them giving the wrong mark, or where the majority of the markers do not agree with the definitive mark as the correct mark), we added the mode of marks obtained at the item level on the seed scripts to arrive at an alternative 'definitive' mark (referred to here as the SIM – Sum of Item Modes) against which the awarded marks could be compared. Note that the SIM did not involve the original definitive marks decided by the senior examining panel.

Figure 10 shows the difference between the sum of item modes and the original definitive mark for each seed script for all the components. The dots connected by a line show the difference (SIM-Definitive mark) according to seed scripts.

The x-axis in Figure 10 shows the sequence number of seed scripts in a component. The scripts are ordered by their total definitive mark, from low to high. The line at 0 on the y-axis shows the line of no difference (complete agreement at whole script level between the SIM and definitive mark). Differences above this line indicate, on average, lenient marking by markers as compared to the definitive mark whereas those below the line indicate severe marking.

The graphs in Figure 10 show that the difference between the SIM and the definitive mark was higher in the Long components in each pair. The scripts with some of the largest differences between the two marks were script #8 in Pair 1, Long component and script #6 in Pair 4, Long component. These scripts were also identified in Figure 7 (which showed differences between awarded and definitive marks). As is evident from Figure 7, most of the markers did not agree with the definitive mark on these scripts and it could be the case that the definitive mark was not the appropriate gold standard mark for the scripts.

Figure 10 shows that the SIM and definitive marks were almost the same for all the Short components. The differences in all the seed scripts were limited to the -2 to +2 range. For the Long components, about 76% of the seed scripts across all the components had the differences between -3 to +3 range.

The plots given in Figure 10 had a similar pattern to the connecting lines of average (mean) differences between the awarded and the definitive marks in Figure 7. This suggests that the SIM was an appropriate average of awarded marks for this analysis.

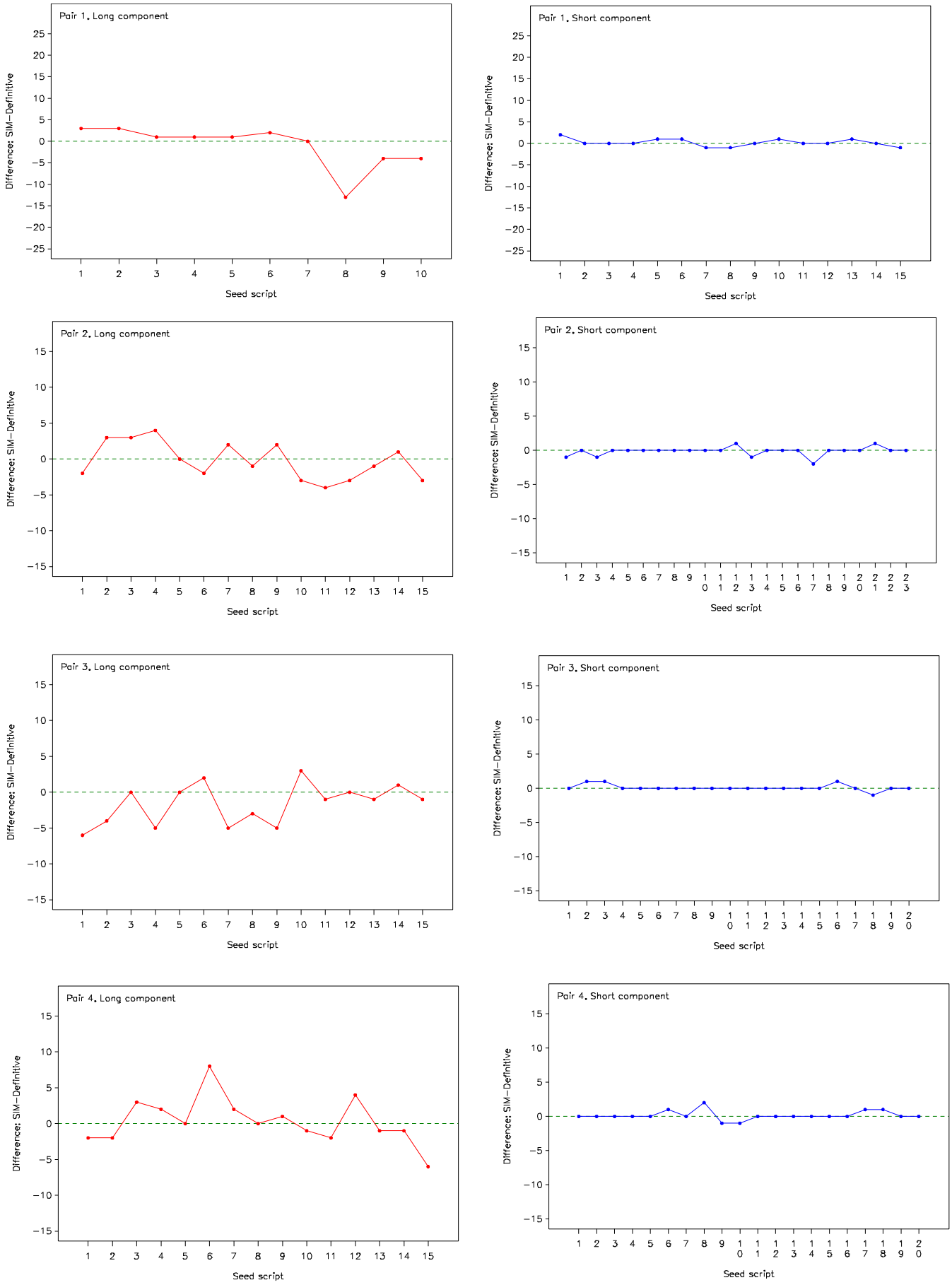


Figure 10: Differences between SIM (Sum of Item Modes) and definitive mark, displayed for each seed script.

Table 4.5 gives the mean and standard deviation of the differences between the awarded marks and the SIM. For comparison it also gives the mean and standard deviation of the differences between awarded marks and the original definitive marks (also presented in Table 4.2).

Table 4.5: Summary distribution of differences of Awarded-SIM and Awarded-Definitive marks

Pair Num	Type	Comp. Label	Paper Total	Mean Awarded-SIM	SD Awarded-SIM	Mean Awarded-Definitive	SD Awarded-Definitive
1	Long	1L	90	1.39	6.48	0.29	7.43
1	Short	1S	90	-0.16	1.97	-0.09	1.97
2	Long	2L	60	0.17	3.13	-0.14	3.68
2	Short	2S	60	0.10	1.13	-0.02	1.17
3	Long	3L	60	0.85	3.12	-0.78	3.51
3	Short	3S	60	-0.12	0.71	-0.01	0.75
4	Long	4L	60	0.26	3.19	0.59	3.84
4	Short	4S	60	0.10	1.18	0.24	1.13

Table 4.5 shows that the mean differences of the SIM from the awarded marks were similar to the mean differences of the definitive marks from the awarded marks in the Short components. There was more variation in the mean in the Long components where the mean difference was higher when the SIM was used (except in component 4L). A comparison of the standard deviations shows that the spread of differences was very similar for all the Short components. The Long components tended to have a slightly narrower spread of differences between the awarded and the SIM compared to the differences between the awarded and the definitive marks. This is to be expected because using the SIM 'takes out' the contribution of systematic differences between the SIM and the definitive mark across the seed scripts.

4.5 Item-level agreement

The focus of this report was to investigate marker agreement at the script level. However, it would be worthwhile here to have a brief overview of agreement at the item level as well.

Table 4.6 gives the number of seed item marking events for each component. This value gives the number of items in the paper multiplied by the marking events at the seed script level. The table also gives the number and the percentage of the item marking events where the awarded mark was exactly equal to the definitive mark.

Table 4.6: Agreement between awarded and definitive mark at item level

Pair Num	Type	Comp. Label	Qualification	# Item marking events	# Exact agreement events	% Exact agreement
1	Long	1L	GCE	4464	835	18.7
1	Short	1S	GCE	9840	6599	67.1
2	Long	2L	GCE	8268	1452	17.6
2	Short	2S	GCE	13560	10236	75.5
3	Long	3L	GCSE Unit	2755	1552	56.3
3	Short	3S	GCSE Unit	6256	5396	86.3
4	Long	4L	GCSE Unit	6650	2659	40.0
4	Short	4S	GCSE Unit	25499	20758	81.4

Table 4.6 shows that, in each pair, the percentage of items having an exact agreement was considerably higher in the Short component than the Long component. This was not a surprising finding given the different type of items in the two categories of components. The average exact-agreement percentage was 33.2% for the Long components and 77.6 % for the Short components.

Figure 11 gives the average marking accuracy percentage of items according to their maximum mark across all the four pairs.

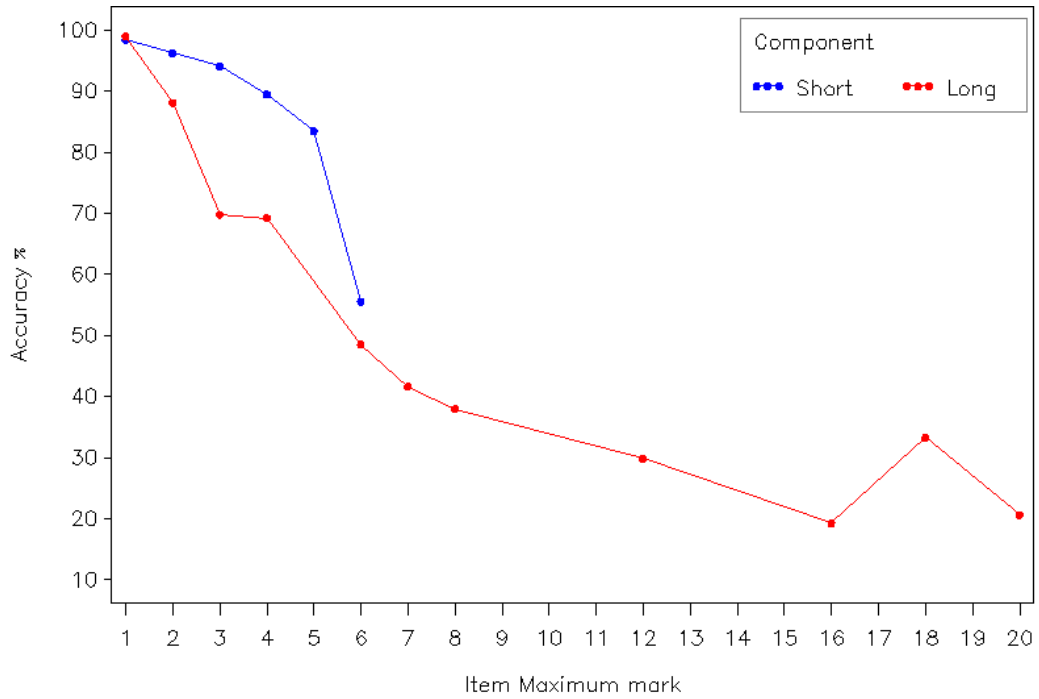


Figure 11: Average marking accuracy percentage against maximum numeric mark of items.

Figure 11 shows that all the items in the Short components were less than or equal to six marks each whereas this limit was 20 marks in the Long components. The figure shows that the lower the maximum mark of an item, the higher the average accuracy percentage. The average accuracy percentage was higher in the Short components for all except 1-mark items where average accuracy was similar for both the Long as well as the Short components. Items having a maximum mark of six or below had an average accuracy percentage from 50% to 100%. Items worth 12 or more marks had lower spread of average accuracy, which was more or less below 30%. This indicated that the items which were worth less (and were more likely to be objective or short-answer type questions) had a higher probability of exact agreement with the definitive mark. Similar results were reported in Bramley (2008) and Raikes and Massey (2007) in which items having a higher maximum mark were found to be associated with lower marker agreement. Bramley (2008) mentioned that the maximum mark of items might capture most of the predictable variation for estimating marker agreement and is likely to be related to the complexity of the cognitive processing tasks which markers need to accomplish to mark the items. A strong relationship between the complexity of the cognitive marking strategy that items require and the relative marking accuracy was also reported in Suto and Nádas (2008, 2009).

5. Effect on classification of examinees

Bramley & Dhawan (2010) showed how a crude indicator of classification consistency could be derived using the SEM calculated from Cronbach's Alpha (referred to as $SEM_{internal}$ in this section). This classification consistency was interpreted as the estimated proportion of examinees who would obtain the same grade on a parallel test. An even cruder indicator of classification consistency can also be derived by treating the standard deviation of the (signed) marker differences as an estimate of the SEM attributable to markers in each component (referred to as SEM_{marker}). This can be interpreted as the estimated proportion of examinees who would obtain the same grade with a different marker.

These two indicators of classification consistency were calculated for each member of each pair of components. An example of the estimated percentage of examinees classified consistently in one of the components (4S) is shown in Table 5.1. The table shows the grade boundaries and the bandwidth (number of marks) available for each grade in this component. The total number and percentage of examinees is also given according to the grades received. The first row in the table (Grade=All) gives the same information for the whole assessment. (The first row has a grade bandwidth of 61 because the maximum mark for the component was 60, giving 61 possible scores on the test, including zero). The last two columns give the comparison of the estimated percentage of examinees with a given grade who were likely to get the same grade, using test-related and marker-related sources of error.

Table 5.1: Example of estimated classification consistency, component 4S

Grade	Grade boundaries	Grade bandwidth (marks)	Number of examinees	% of examinees	Estimated % consistently classified (test)	Estimated % consistently classified (marker)
All		61	11428	100.0	62.9	86.6
A*	46	15	1143	10.0	79.9	90.2
A	38	8	2034	17.8	65.0	87.3
B	30	8	3067	26.9	65.1	87.5
C	23	7	2896	25.4	61.0	86.7
D	18	5	1417	12.4	49.3	82.9
E	15	3	436	3.8	32.4	71.1
U	0	15	435	3.8	80.6	94.7

Table 5.1 shows that, in this component, the proportion of examinees consistently classified across each grade was higher when SEM_{marker} was used. The first row gives the aggregate difference for this component in the estimated percentage of candidates who would get the same grade using $SEM_{internal}$ (62.9%) and SEM_{marker} (86.6%). The aggregate differences (similar to the first row of Table 5.1) for all the eight components are shown in Table 5.2. The table also shows the comparison of the SEM and the Bandwidth:SEM ratio according to the two sources of error, test-related and marker-related.

Table 5.2: Estimated classification consistency for all the selected components

Pair Num	Comp. Label	Grade Bandwidth	$SEM_{internal}$	SEM_{marker}	Bandwidth: SEM (test)	Bandwidth: SEM (marker)	Est. % consistently classified (test)	Est. % consistently classified (marker)
1	1L	7	6.05	7.43	1.16	0.94	49.7	44.5
1	1S	7	4.92	1.97	1.42	3.55	59.7	82.0
2	2L	4	4.77	3.68	0.84	1.09	50.0	56.2
2	2S	5	3.75	1.17	1.33	4.28	58.4	83.8
3	3L	6	3.73	3.51	1.61	1.71	52.1	54.2
3	3S	6	3.28	0.75	1.83	7.96	58.5	88.7
4	4L	7	4.72	3.84	1.48	1.82	47.4	54.5
4	4S	8	3.49	1.13	2.29	7.10	62.9	86.6

Table 5.2 shows that the SEM_{marker} was lower than the $SEM_{internal}$ for all components except 1L where the reverse was true. The difference between the two SEM values was higher for the Short component in each pair. Similar results were obtained for the Bandwidth:SEM ratio where the values derived from SEM_{marker} were higher except in the component 1L.

The last two columns of Table 5.2 show the comparison of classification consistency. The percentages using $SEM_{internal}$ were not vastly different from those reported in Bramley & Dhawan (ibid.) in which the components used mainly consisted of short-answer or objective-type questions. In Table 5.2, the values given in the last column (using SEM_{marker}) were higher than those given in its previous column (using $SEM_{internal}$) for all the components except 1L. The comparison is also shown in Figure 12. The diagonal line in the graph is an identity line. The values below this line represent components in which the estimated proportion of consistently classified examinees was higher using SEM_{marker}.

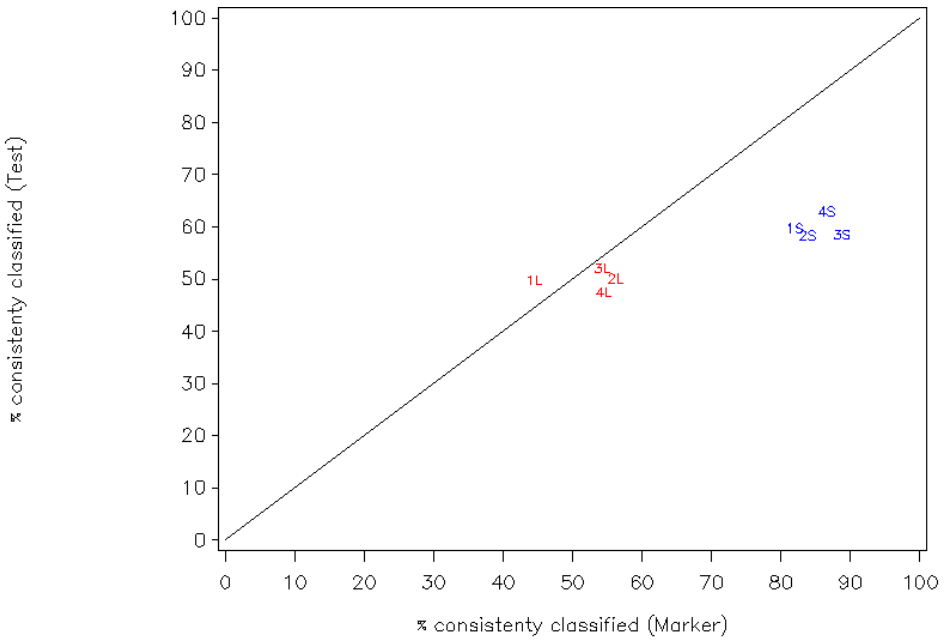


Figure 12: Classification consistency using $SEM_{internal}$ vs. SEM_{marker}.

In Figure 12, all the points (except 1L) are below the diagonal line which shows that the possibility of examinees getting the same grade was higher using SEM_{marker} than $SEM_{internal}$. The figure also highlights that classification consistency was higher for SEM_{marker} in the Short components than the Long components.

In this section, the results were compared using crude indicators of $SEM_{internal}$ and SEM_{marker}. The $SEM_{internal}$ was calculated using all the examinees who sat the paper whereas the SEM_{marker} was estimated using only the seeding scripts which therefore might not be as accurate as $SEM_{internal}$. Also, this analysis did not completely segregate the two sources of error. Hutchison & Benton (2009, p40-41) point out that any estimate of Cronbach's Alpha will take into account a certain proportion of variation attributed to markers. A full generalizability analysis could in principle handle the various sources of error in a more rigorous framework (Johnson & Johnson, 2009), but it is unlikely that the kind of data available from seeding scripts would be comprehensive enough to allow such an analysis.

6. Discussion

In this study higher marker variability was observed in the components having long-answer or essay-type questions than in the components having objective or short-answer questions. Overall the markers were neither too severe nor too lenient. The spread of marker differences was larger in the Long component than its corresponding Short component in all the four pairs. The estimates of internal consistency were better for the Short components. All the components selected in this study were *single* units/components of larger assessments. Overall (composite) reliability of the whole assessment is likely to be higher as shown in Bramley & Dhawan (2010).

The crude analyses of classification consistency showed that a higher proportion of examinees would be consistently classified on the same test with a different marker than on a parallel test, but the difference between the two scenarios was much less for the Long components than the Short ones. In other words, examinees in the Short components would be more likely to get the same grade if their work was marked by a different marker than if they sat a different (parallel) test, but this was less clearly the case for the Long components.

The results were not too unexpected as greater amount of deviation from the definitive mark is likely to occur where markers have to mark extended response questions by applying complex mark schemes. Massey & Raikes (2006) found that there was more agreement on objective items than on points-based and levels-based items which have more complex mark schemes[8]. Similar results were reported in Bramley (2008), Suto and Nádas (2008) and Black et al. (2010). These studies found that the more constrained the item, the higher was the level of marker agreement. The less constrained items usually tend to have more complex mark schemes for which markers have to apply their judgement and interpretation to a greater degree. For a discussion of the cognitive strategies involved in the marking of examinations and what goes through an examiner's mind see Suto & Greatorex (2008). Table A2 in the appendix gives a comparison of the relative complexity of mark schemes between less constrained and more constrained questions. The mark schemes of the three questions given in the table were extracted from the same component (a unit of GCE Business Studies) from the June 2011 session.

Setting the same tolerance for deviations from the exact definitive mark of a seed script would not be fair to markers in components having a higher proportion of essay-type questions. This is already recognised because the Long components in each pair did have higher tolerances – however, this research has suggested that either they need to be higher still, or that the marking of the Long components was not always within acceptable limits. The method of deriving a tolerance value assumes significance in ensuring optimal use of the value for marker monitoring. The most straightforward method of deciding tolerance at the script level is to calculate it as a certain percentage of the paper total, but this does not allow for differences in question type. A slightly more complex but useful method would be to calculate tolerance values for each item and then arrive at script-level tolerance. This would allow more fine-grained information to be fed into the process of marker monitoring. Other factors like length of the answers required and complexity of the mark scheme could also be taken into consideration to arrive at an appropriate value. If the tolerance value is too small, it is not fair to the markers. On the other hand, setting a large tolerance value might lead to ineffective monitoring of markers. Therefore, setting tolerance at the right level, particularly for essay-type questions, would be an important step in effective and fair monitoring of markers.

In this study attempts were made to include a wide range of assessments. However, key subjects containing essay-type questions like English and History could not be included either because they were not marked on-screen or did not fulfil the criteria used for matching the Long and the Short components. More and more components are likely to be marked on-screen in

[8] See appendix Table A1 for definition of objective, points-based and levels-based items.

the future which should allow analysis of a wider range of components for investigating inter-marker reliability. The use of on-screen marking has made the process of gathering evidence of performance of markers considerably easier and increasingly informative. While even more fine grained results could be obtained using complex methods in IRT or Generalizability Theory, the cause of examinees would be better served by focussing on qualitative mechanisms to reduce the gap between the awarded mark and the deserved mark. Black, Suto & Bramley (ibid.) introduced a comprehensive framework of actions that could potentially improve marking quality using features of questions, mark schemes and examinee responses. The authors also discussed the potential impacts of these actions on validity, and their cost effectiveness. Future research in inter-marker reliability should take into consideration the relative complexity of mark schemes and the cognitive demands placed on the markers while marking different types of questions.

It should be noted that the data used in this study came from live on-screen monitoring of markers on seeding scripts which represent a very small percentage of the total number of scripts marked in a live examination session. The statistics obtained from monitoring markers using seeding scripts is only a part of the larger process of quality control. It does not take into account other quality control measures such as re-marking the scripts of markers who were stopped from marking because their work was deemed to be unsatisfactory. Therefore, some of the data described in this report would trigger quality assurance processes, resulting in appropriate action being taken, so that the results the candidates receive are as accurate as possible. The schools/examinees have the option to make an ‘enquiry about results’ if they are unhappy with the results which may lead to a re-mark. In extreme cases where disputes cannot be resolved, there is an official appeals procedure. The aim of the whole process of quality control, which includes marker monitoring, is to ensure that examinees are awarded the final outcome as accurately as possible based on their performance in the examination.

References

Benton, T. (2011). *Empirically optimising an active marking quality control system.* NfER talk given at Cambridge Assessment, Cambridge, November 2011. Available at http://www.assessnet.org.uk/e-learning/mod/resource/view.php?id=2424. Accessed 5/12/2011.

Black, B., Suto, W.M.I, & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice, 18(3),* 295-318.

Black, B., Curcin, M. & Dhawan, V. (2010). *Investigating seeding items used for monitoring on-line marking: factors affecting marker agreement with the gold standard marks.* Paper presented at the annual conference of the International Association for Educational Assessment (IAEA), Bangkok, Thailand, August 2010.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication, 4*, 22-28.

Bramley, T. (2008). *Mark scheme features associated with different levels of marker agreement.* Paper presented at the British Educational Research Association (BERA) annual conference, Heriot-Watt University, Edinburgh, September 2008.
http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers
Accessed 30/11/2011.

Bramley, T., & Dhawan, V. (2010). *Estimates of reliability of qualifications*. Coventry: Ofqual.
http://www.ofqual.gov.uk/files/reliability/11-03-16-Estimates-of-Reliability-of-qualifications.pdf
Accessed 16/11/2011.

Hutchison, D., & Benton, T. (2009). *Parallel universes and parallel measures: estimating the reliability of test results.* Slough: NfER. Ofqual/10/4709 http://www.ofqual.gov.uk/files/2010-02-01-parallel-universes-and-parallel-measures.pdf Accessed 30/11/2011.

Johnson, S., & Johnson, R. (2009). *Conceptualising and interpreting reliability*. Assessment Europe. Ofqual/10/4709.

Massey, A.J. & Raikes, N. (2006). *Item level examiner agreement.* Paper presented at the 2006 Annual Conference of the British Educational Research Association, 6-9 September 2006, University of Warwick, UK.

Raikes N. & Massey, A. J. (2007). Item-level examiner agreement. *Research Matters: A Cambridge Assessment Publication, 4,* 34–37.

Suto, W.M.I. & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal, 34(2),* 213-233.

Suto, W.M.I. & Nádas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education, 23(4),* 477-497.

Suto, W.M.I. & Nádas, R. (2009). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education, 24(3)*, 335-377.

Appendix

Table A1: Classification of items according to the kind of marking required, as given in Raikes and Massey (2007).

Objective marking – items that are objectively marked require very brief responses and greatly constrain how candidates must respond. Examples include items requiring candidates to make a selection (e.g. multiple choice items), or to sequence given information, or to match given information according to some given criteria, or to locate or identify a piece of information (e.g. by marking a feature on a given diagram), or to write a single word or give a single numerical answer. The hallmark of objective items is that all credit-worthy responses can be sufficiently pre-determined to form a mark scheme that removes all but the most superficial of judgements from the marker.

Points based marking – these items generally require brief responses ranging in length from a few words to one or two paragraphs, or a diagram or graph, etc. The key feature is that the salient points of all or most credit-worthy responses may be predetermined to form a largely prescriptive mark scheme, but one that leaves markers to locate the relevant elements and identify all variations that deserve credit. There is generally a one-to-one correspondence between salient points and marks.

Levels based marking – often these items require longer answers, ranging from one or two paragraphs to multi-page essays or other extended responses. The mark scheme describes a number of levels of response, each of which is associated with a band of one or more marks. Examiners apply a principle of best fit when deciding the mark for a response.

Table A2: Example mark schemes

Question			Expected Answers	Mks	Rationale/Additional Guidance
1[9]			**Assess the likely impact on the motivation of JKL's employees of the proposal to increase capacity utilisation in each childcare centre.** JKL appears to be a business which respects its employees, is keen to look after them by paying a salary above the industry norm and thus ensure they are motivated and happy. One of Harriet's stated aims is to raise the standards and status of the childcare profession. One potential impact of the proposal to increase capacity utilisation above 100% is on the staff-child ratio. We are told that this is currently above the required standard so that the workers will have more time to play with, supervise and look after the children in their care. If we assume that the workers are currently motivated then any increased pressures on their work may reduce this motivation. This may also have implications in terms of safety standards and the level of customer happiness. This is not something to be jeopardised - given the current concerns about increased competition and the imminent opening of a new centre. Increased pressures on workers may also make it more difficult for managing staff rotas which may then have an impact on the current availability of flexible rotas. As most employees will be women with children themselves and working part-time, there may be serious problems. Up to 40% of the staff are currently working towards some form of childcare qualification. In addition, JKL provides paid study leave and allows training time for at least 10 in-house training courses. Training, and the opportunity for promotion which comes with it, are seen as motivators by Herzberg and Maslow. Increased pressure		

[9] The question numbers have been changed in this report.

Question			Expected Answers	Mks	Rationale/Additional Guidance
			on work will jeopardise this and may ultimately mean that JKL's employees are less well qualified and so do their job less well. (The proposal to introduce the Quality Counts initiative may also be affected by increased capacity utilisation.) Imposing any changes may also cause problems as it contradicts the theory of Mayo, with regards to the human relations approach of workers feeling involved and appreciated - something present at the moment. Ultimately, JKL need to consider the balance between increasing the efficient use of its factors and the impact which this has on the workforce, in an industry which is highly reliant on the motivation, quality and dedication of its staff. Will this increase in capacity utilisation be a short-run issue or may it jeopardise the long-term aims of the business? **Level 4** Some evaluation shown by taking a judgemental approach to the consequences for JKL of changes in its workers motivation. **(16-12)** **Level 3** Some analysis of the possible impact on the motivation of JKL's employees of increased workloads linked to higher capacity utilisation. **(11-7)** **Level 2** Some understanding shown of how motivation may be affected by changes in capacity utilisation. **(6-3)** **Level 1** Some knowledge about factors affecting motivation and/or motivational theories. **(2-1)**	**[16]**	**<u>Default marks</u>** L4 – 14 L3 – 9 L2 – 5 L1 – 2

Question			Expected Answers	Mks	Rationale/Additional Guidance
2			**Identify two leadership styles.** One mark for each of the correct identifications. Any **two** from: • autocratic • democratic • paternalistic • laissez-faire (allow vaguely correct spellings)	**[2]**	Accept other recognisable leadership style, such as bureaucratic, technocratic.
3			**A business is considering purchasing a new piece of machinery at a cost of £50,000.** **The machinery is expected to last for five years and produce annual net cash inflows as follows.** Year Net cash inflow (£) 0 (50,000) 1 20,000 2 30,000 3 40,000 4 40,000 5 40,000 **Calculate the Accounting Rate of Return (ARR) for the machinery.** **Total Profit = £170,000 - £50,000 = £120,000 [1]** **Average Annual Profit = £120,000 / 5 = £24,000 [1]** **ARR = £24,000 x 100 = 48% [2]** **£50,000** **OFR**	**[4]**	There is an alternative method to calculating ARR found in some books which has to be accepted. £170,000/5 = £34,000 £34,000/£50,000 = 68% Award four marks for correct answer (even with no working). Award three marks for 0.48 (OFR). Look for the number of mistakes made by a candidate. This can help in the marking of more complicated attempts.

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2013

© Crown copyright 2013

You may re-use this publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence. To view this licence, visit The National Archives; or write to the Information Policy Team, The National Archives, Kew, Richmond, Surrey, TW9 4DU; or email: psi@nationalarchives.gsi.gov.uk

This publication is also available on our website at www.ofqual.gov.uk

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

Office of Qualifications and Examinations Regulation
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

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

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