Marking Reliability of the Key Stage 2 National Curriculum English Writing Tests in England

February 2014

Ofqual/14/5377

Contents

Summary

In England, pupils at the age of 11 take the Key Stage 2 National Curriculum English test annually. Until 2011, the English test was composed of a reading component and a writing component. Pupils were awarded a National Curriculum attainment level for the overall English and for each component based on their performance in the test.

The English writing test assesses pupils' ability in four main areas: sentence structure and punctuation; text structure and organisation; composition and effect; and spelling. Pupils' scripts are marked by human markers. Markers are trained and standardised before live marking to ensure marking consistency. The quality of marking of the markers is further monitored through the embedding of benchmark scripts in live marking.

The present study investigates marking reliability of the Key Stage 2 English writing tests administered in 2010 and 2011 through the analysis of data collected from the marker standardisation, benchmarking and standards setting processes. The uncertainties in pupils' writing test scores associated with inconsistency in marking between markers are explored. The effects of factors such as the nature of the tasks and the quality of pupils' work on marking consistency and accuracy are examined. There was considerable variability in the marking of the Key Stage 2 English writing tests that were investigated, reflecting the nature of the tests. Findings are presented using an intuitive and informative manner for practical applications.

1 Introduction

1.1 End of Key Stage 2 National Curriculum tests in England

The school curriculum in England for pupils aged 5–16 is structured into four key stages as a result of the introduction of the National Curriculum in the late 1980s (Qualifications and Curriculum Authority (QCA), 1999; Department for Education (DfE), 2011). Each key stage covers a number of years of schooling: Key Stage 1 is for years 1–2 (ages 6–7), Key Stage 2 is for years 3–6 (ages 7–11), Key Stage 3 is for years 7–9 (ages 11–14), and Key Stage 4 is for years 10–11 (ages 15–16). The National Curriculum sets out the programmes of study for individual subjects that schools need to follow, and the attainment targets that pupils are expected to achieve at the end of each key stage. The core subjects, which include English, mathematics and science, are compulsory for schools across Key Stages 1–4. The foundation subjects, which include art and design, design and technology, geography, history, information and communication technology, and others, vary at different key stages (DfE, 2011).

At the end of each key stage, pupils are assessed for their performance in the National Curriculum through both teacher assessment and external tests. The end of Key Stage 2 National Curriculum tests include tests in English (reading and writing), mathematics and science. The English and mathematics tests are taken by the whole national cohort, whereas the science tests are taken by pupils from a nationally representative sample of schools every year (QCA, 1999; Whetton, 2009; Isaacs 2010; He et al., 2012a). The development of the Key Stage 2 national tests is carried out using standard test development procedures. Tasks and items are created by experienced assessment experts and evaluated by review panels consisting of experts from a variety of relevant areas – including curriculum subjects, inclusion and cultural awareness – and from different perspectives – such as teachers, local authorities and potential markers. Initially selected tasks and items are used to construct tests for pre-testing before they are used in live testing. The pre-testing process is well-defined and rigorous. The purposes of the pre-testing are to evaluate the quality of tasks and items further, to produce item statistics, to ensure that the tests are at the appropriate difficulty level for the target population, and to produce initial performance level boundary scores by equating the tests with anchor tests. Test equating is done to ensure that the comparability of performance standards over time is maintained for individual subjects, which is crucial for monitoring national performance over time.

Pupils taking the Key Stage 2 National Curriculum tests are awarded a National Curriculum attainment level at both component and subject level for English, and at subject level for mathematics and science based on their test scores. A standard-setting process (QCA, 2009), which involves the use of both statistical information and professional judgement of the quality of sampled pupils' work, is used to set

thresholds for National Curriculum performance at levels 3, 4 and 5 for the mark distributions of the tests. In addition, outcomes from the tests are aggregated and published nationally in order to monitor national attainment or evaluate school performance in the individual subjects at the end of the primary phase.

1.2 The Key Stage 2 English writing tests

The National Curriculum Key Stage 2 English test is designed to assess a range of pupils' reading and writing skills. The test has two components: a reading test and a writing test. Both tests are taken under exam conditions. The maximum available mark on each of the components is 50.

The writing test focuses on assessing pupils' ability to:

- write imaginative, interesting and thoughtful texts;
- produce texts that are appropriate to task, reader and purpose;
- organise and present whole texts effectively, sequencing and structuring information, ideas and events;
- construct paragraphs and use cohesion within and between paragraphs;
- vary sentences for clarity, purpose and effect;
- write with technical accuracy of syntax and punctuation in phrases, clauses and sentences;
- select appropriate and effective vocabulary (this is not assessed separately, but contributes to text structure and organisation, and composition and effect);
- use correct spelling (assessed through the spelling test).

(List adapted from Qualifications and Curriculum Development Agency (QCDA), 2011.)

The writing test is composed of a short task (worth a maximum of 12 marks), a long task (worth a maximum of 31 marks) and a spelling test section (worth 7 marks). To facilitate marking of the two tasks, the writing skills that are assessed by the tasks are organised into four mark scheme strands:

- sentence structure and punctuation
- text structure and organisation
- composition and effect
- handwriting.

Table 1 shows how the marks are allocated between the different mark scheme strands. For both tasks, the composition and effect (CE) marking strand has the highest maximum available marks among the different strands (8 for the short task, and 12 for the long task, respectively).

Table 1 Number of marks allocated to different mark scheme strands for the 2010 and 2011 Key Stage 2 English writing tests (excluding the spelling test section)

Skills	Short task	Long task
Sentence structure, punctuation and text organisation (SSPTO)	4	
Sentence structure and punctuation (SSP)		8
Text structure and organisation (TSO)		8
Composition and effect (CE)	8	12
Handwriting (HW)		3
Total	43	

A detailed mark scheme has been developed for markers to use for evaluating pupils' work. For each mark scheme strand, the allocated maximum mark is divided into a number of bands, and detailed marking criteria for each band are provided so that markers will be able to award appropriate marks to pupils' responses. The criteria specify typical characteristics of pupils' work in different mark bands, and annotations on example scripts illustrate how to look for features in the writing and how to weigh the features to reach an appropriate mark.

1.3 Studies of reliability of marking in England

Reliability refers to the consistency of assessment results over replications of the assessment procedure (Brennan, 2001; Haertel, 2006). Specifically, marking reliability refers to the extent to which marks from different markers on the same piece of work from a test-taker are consistent. Reliability is a prerequisite for validity – if the same piece of work is marked by a different marker and a substantially different result is produced, then the validity of any inferences based on the results from the test would be difficult to establish. Test results should therefore have the necessary degree of reliability for it to be possible that inferences made, based on them, might be valid.

The quality of marking – particularly the reliability of marking – of assessments for subjects for both national tests and public exams in England that involve human

markers has been a concern for many stakeholders (see, for example, House of Commons Children, Schools and Families Committee, 2008; Ward, 2009; Bew, 2011; Murray, 2011; Headmasters' and Headmistresses' Conference, 2012; Ofqual, 2012). There has been considerable research into the reliability of marking for the now defunct Key Stage 3 National Curriculum tests and public exams such as GCSEs (General Certificate of Secondary Education exams taken by students at the age of 16 in England); see, for example, Murphy, 1982; University of Cambridge Local Examinations Syndicate, 2002; Baird et al., 2004; Baker et al., 2008; Royal-Dawson, 2005; Royal-Dawson and Baird, 2009; Royal-Dawson et al., 2008; Newton, 2009; Fowles, 2009; Black et al., 2011; Suto et al., 2011; Bramley and Dhawan, 2012; Dhawan and Bramley, 2012; Johnson et al., 2010; Johnson and Johnson, 2012; Baird et al., 2012; He and Opposs, 2012; Opposs and He, 2012. However, little is understood about the level of reliability of marking of the Key Stage 2 National Curriculum tests (Benton, 2006; Newton, 2009), particularly the Key Stage 2 English reading and writing tests.

1.4 Aims of the present study

The present study is intended to investigate the reliability of marking of the Key Stage 2 English writing tests administered in 2010 and 2011, excluding the spelling test section. Specifically, this study focuses on the following aspects of marking reliability:

- the level of unreliability in marking of the Key Stage 2 National Curriculum English writing tests;
- the effect of the quality of pupils' work on marking reliability at both question paper or script level and individual marking strand level;
- variability in marking between markers over time.

This study also compares marker-related standard error of measurement and test-related standard error of measurement in test scores.

2 Methodology

2.1 Data collection

Data collected from marker standardisation, benchmarking and standards setting for the Key Stage 2 English writing tests administered in 2010 and 2011 were used in this study.

The majority of markers have primary school teaching experience. All markers receive comprehensive training, at a regional meeting, before they start marking pupils' work, the aim being to train the markers to apply the mark scheme consistently at the agreed standard when assessing pupils' work. This involves discussion and analysis of the marking criteria and annotated exemplars of pupils' work so that the markers can reach consensus on the application of the marking criteria to specific features of pupils' work.

Following marker training, two quality assurance processes, referred to as standardisation and benchmarking, are employed to ensure that a high quality of marking is achieved. These processes involve markers marking sets of scripts that have already had their marks agreed by a group of senior markers. Each marker's decisions are compared with the scripts' 'definitive' marks that have already been determined by the senior markers. These standardisation and benchmarking scripts are selected from the pupils who took part in the pre-testing of the test. There are two rounds each of standardisation and benchmarking.

Standardisation takes place immediately following training and prior to the marker commencing marking. In Standardisation 1, markers mark ten standardisation scripts, and their marks on individual marking strands and the overall scripts are compared with the scripts' definitive marks. Markers with differences between their marks on the scripts and the script definitive marks above a pre-set threshold are asked to mark a further five standardisation scripts, which is referred to as Standardisation 2. Markers who fail Standardisation 2 are not allowed to mark.

To monitor the quality of marking during the subsequent marking period, all markers are required to mark ten further benchmarked scripts after they have marked about a third of the scripts allocated to them. This is referred to as Benchmarking 1. Markers' marks on these benchmarked scripts are compared with the scripts' definitive marks and are used to assess the quality of their marking. Markers who fail Benchmarking 1 are stopped from marking and have their unmarked scripts reassigned to other markers. Scripts that have already been marked by the failed markers are re-marked by other markers. After they have marked about two-thirds of their allocated scripts, all markers are required to mark a further ten scripts, which is referred to as Benchmarking 2. Once again, markers who fail Benchmarking 2 are stopped from marking and all of their scripts are reassigned to other markers.

To ensure that the scripts used for standardisation and benchmarking are reasonably representative of pupils' work in live testing, they contain work of varying quality and hence varying definitive marks.

To assist standards setting, the mark distribution for over 3,600 pupils from a representative sample of schools was generated. For these pupils, detailed marks on individual marking strands were also recorded in order to examine the performance of individual tasks and the overall question paper.

2.2 Data analysis

To achieve the aims set for this study, the data collected from the 2010 and 2011 standardisation, benchmarking and standards setting processes were analysed for:

- variability in marks on the same scripts/tasks/marking strands from different markers;
- variability of the difference between markers' marks and scripts' definitive marks or between markers' marks and markers' mean marks on the scripts;
- the internal consistency reliability of the two tests;
- standard error of measurement (SEM) in test scores.

SPSS and Microsoft Excel were used to analyse the data for basic descriptive statistics and fit statistical models to the data.

Markers who failed the standardisation and benchmarking processes were excluded from the analysis.

3 Results and findings

3.1 Variability of marking at question paper (script) level

The scripts used for standardisation and benchmarking have varying definitive marks; Table 2 lists the range of definitive marks and the mean marks of the scripts.

Table 2 Number of scripts used for and number of markers involved in standardisation and benchmarking for the 2010 and 2011 Key Stage 2 English writing tests

	Number of scripts (script code)	Mark range		Mean mark of scripts		Number of markers	
		2010	2011	2010	2011	2010	2011
Standardisation 1	10 (S1–S10)	15–40	14–38	24.2	23.0	1,968	1,812
Standardisation 2	5 (S11–S15)	14–36	8–36	26.0	21.8	57	22
Benchmarking 1	10 (B1–B10)	14–35	8–42	23.8	22.4	1,517	1,812
Benchmarking 2	10 (B11–B20)	12–29	11–38	22.8	26.4	1,462	1,418

Figure 1 shows the distribution of marks assigned to the same standardisation and benchmarking scripts by different markers for the 2010 and 2011 tests. The left half of the graph is for scripts from the 2010 test (code starting ‘10’, with ‘S’ representing standardisation and ‘B’ representing benchmarking), and the right half is for scripts from the 2011 test (code starting ‘11’). The scripts are arranged in order of increasing definitive marks. The horizontal line inside a box represents the median of the marks from the markers on the same standardisation or benchmarking script; the length of a box represents the interquartile range of the marks from the 25th to the 75th percentile for a script; and the vertical T-lines extended from the top and bottom of a box identify the 5th to 95th percentile range of the marks. It is clear from Figure 1 that there is substantial variation in the marks assigned to the same script by different markers.

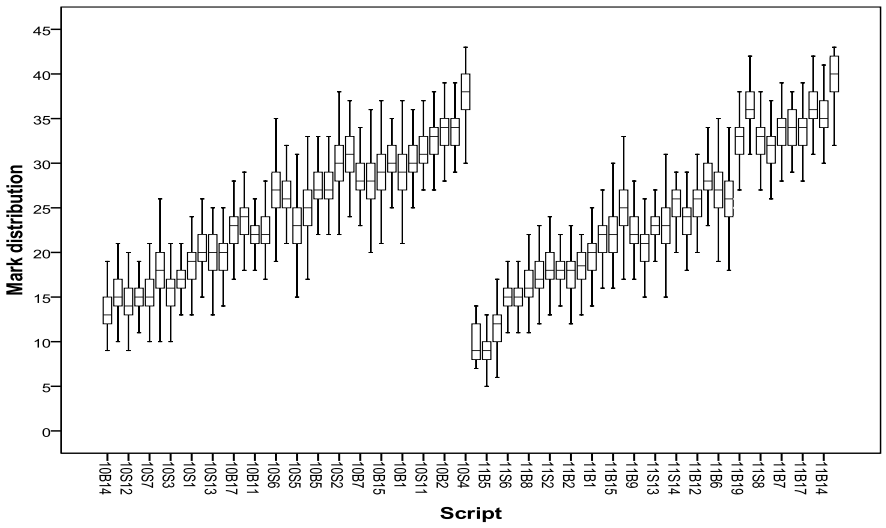

Figure 1 Distributions of marks from different markers assigned to the scripts used for standardisation and benchmarking for the 2010 and 2011 Key Stage 2 English writing tests

Figure 2 depicts the distribution of the difference marks between markers' marks on the standardisation and benchmarking scripts and the scripts' definitive marks. Again the scripts are arranged in order of increasing definitive marks (the same order as in Figure 1). The horizontal line in the middle of the graph at 0 represents the line of no difference between a marker's mark on the script and the script's definitive mark (that is. the mark from the marker is the same as the definitive mark). There is substantial variability in the difference marks for the same script. Figure 2 also suggests that markers' mean marks were generally higher than the scripts' definitive marks for scripts with lower definitive marks (that is the majority of markers' difference marks on these scripts are above the horizontal line at 0) but lower for scripts with higher definitive marks (that is the majority of markers' difference marks on these scripts are below the horizontal line at 0).

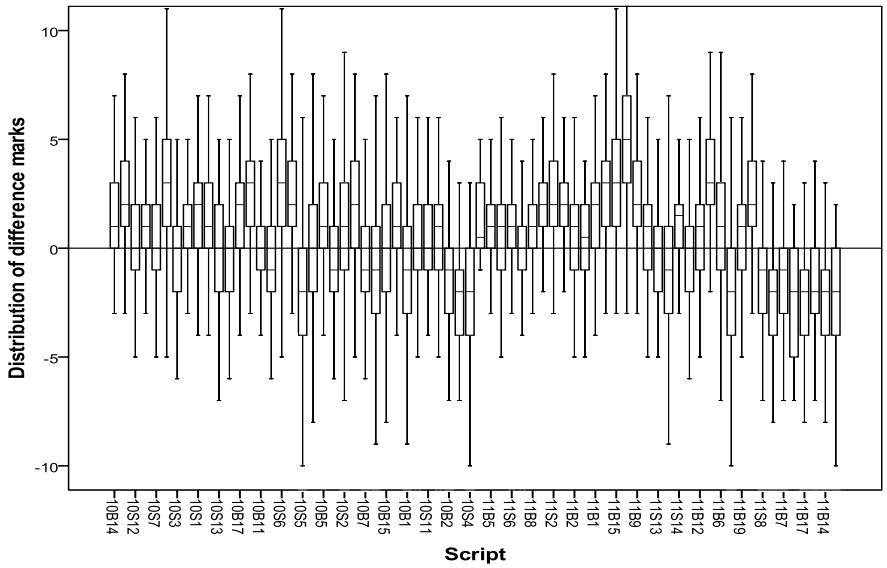

Figure 2 Distributions of difference marks between markers' marks assigned to the standardisation and benchmarking scripts and the scripts' definitive marks for the 2010 and 2011 Key Stage 2 tests

To see the variability in marking between markers in more detail, Figure 3 illustrates the distributions of the difference marks between markers' marks and the script definitive mark on two scripts of different quality: script B4 from Benchmarking 1 of the 2010 test and script B14 from Benchmarking 2 of the 2011 test. Script B4 has a definitive mark of 15, which represents relatively low quality. The mean of the difference marks for the markers on this script is about 3, which suggests substantial 'over-marking' of the script by the markers. In contrast, script B14 has a definitive mark of 38, which represents relatively high quality. The mean of the difference marks for the markers is about −3, which suggests considerable 'under-marking' of the script by the markers. The standard deviation of the difference marks is 2.40 for script B4, which is slightly less than the standard deviation of 2.76 for script B14. The distributions of the difference marks from the markers on the two scripts are relatively symmetric around the mean difference mark. The percentage of markers who had marks in perfect agreement with the script definitive mark (that is when the difference mark between the marker's mark and the script definitive mark is 0) was about 8.3 per cent for script B4 and 11.2 per cent for script B14. And the percentage of markers who had marks that were 1 mark away from the definitive mark (that is when the difference mark is either 1 or −1) was about 16.7 per cent for script B4 and 18.7 per cent for script B14.

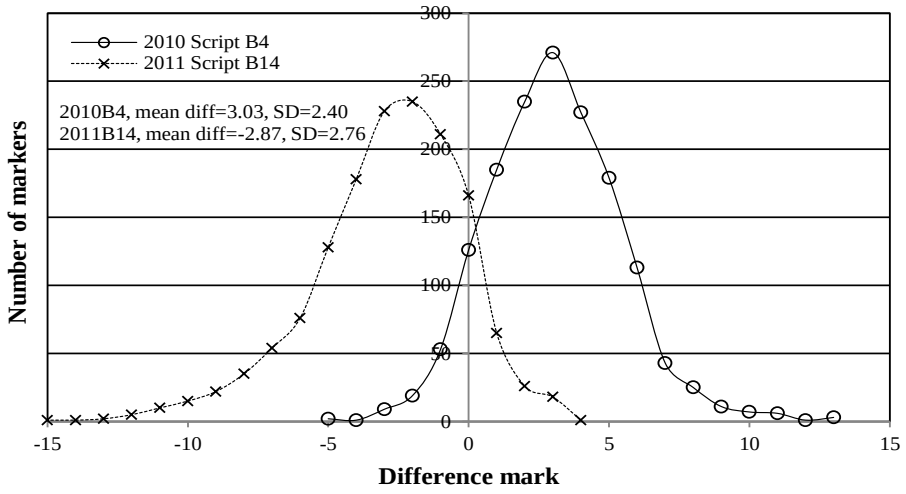

Figure 3 Distributions of difference marks between markers' marks assigned to two benchmarking scripts and the scripts' definitive marks

Figure 4 shows the distribution of the average difference marks between markers' marks on the standardisation and benchmarking scripts, and the scripts' definitive marks. It indicates that markers' mean difference marks are generally positive for lower quality scripts (i.e. scripts with lower definitive marks) but negative for higher quality scripts (i.e. scripts with higher definitive marks), reflecting the effect of regression to the mean. The substantial difference between markers' mean marks and scripts' definitive marks (which were determined by a small group of senior markers) for some of the standardisation and benchmarking scripts would need to be taken into account when exploring the inter-marker reliability of the test, given that the number of normal markers involved in live marking is generally considerably larger than the number of senior markers, and that it is the inter-marker inconsistency in marking between the normal markers that will affect the reliability of the final test scores (see also later discussion).

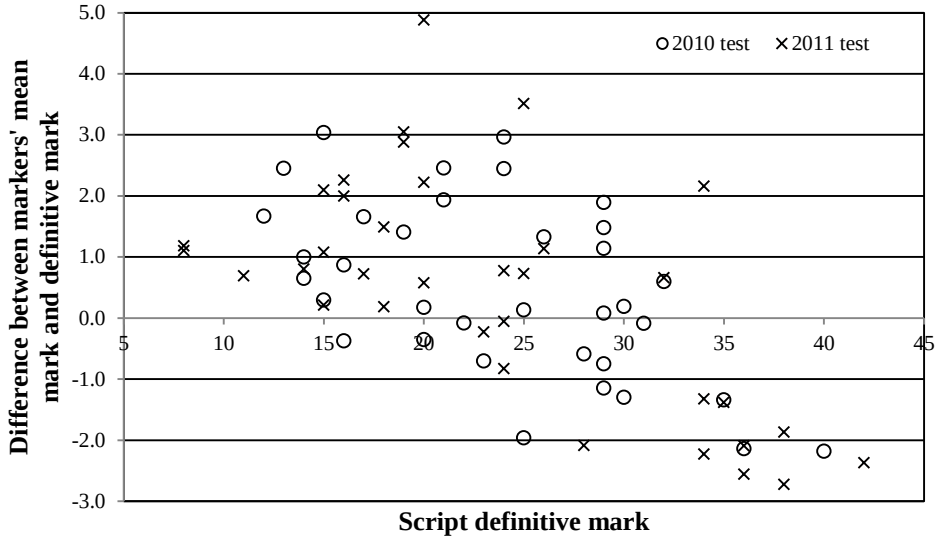

Figure 4 Variation of the difference marks between markers' mean marks on the standardisation and benchmarking scripts, and scripts' definitive marks, with scripts' definitive marks

To illustrate how the quality of pupils' work as reflected by the script definitive marks would affect the consistency of marking between markers, Figure 5 shows how the standard deviations (SDs) of marks assigned to the individual scripts by different markers vary with script definitive mark. The distributions of the standard deviations for the 2010 and 2011 tests are similar. There is a trend that smaller variations in the marks are associated with both low and high quality scripts, while larger variations are associated with scripts of medium quality. This might be expected as it might be easier for markers to reach consensus for both low and high quality scripts, and to apply the marking criteria more consistently, than for scripts of medium quality. Figure 5 also indicates that there is substantial variation in the distributions of marks for scripts with similar definitive marks. For example, there are six scripts from the 2010 test that have a definitive mark of 29 (scripts S2, S15, B7, B13, B15 and B18), and the standard deviation of marks from the markers varies from 2.04 for script B18 to 3.21 for script S15. This variation in marks for scripts with similar definitive marks is likely to reflect the differences in the marks gained by different marking strands between the scripts. Although the marks at the overall script level may be similar for the scripts, the marks on individual mark strands can be considerably different, and the variability of marking between markers for individual mark strand is likely to be affected by the quality of pupils' work. A detailed inspection of the marks gained by individual marking strands for scripts S15 and B18 indicated that the strand marks are more extreme for script B18 than for script S15, which could have resulted in script B18 being more consistently marked than script S15.

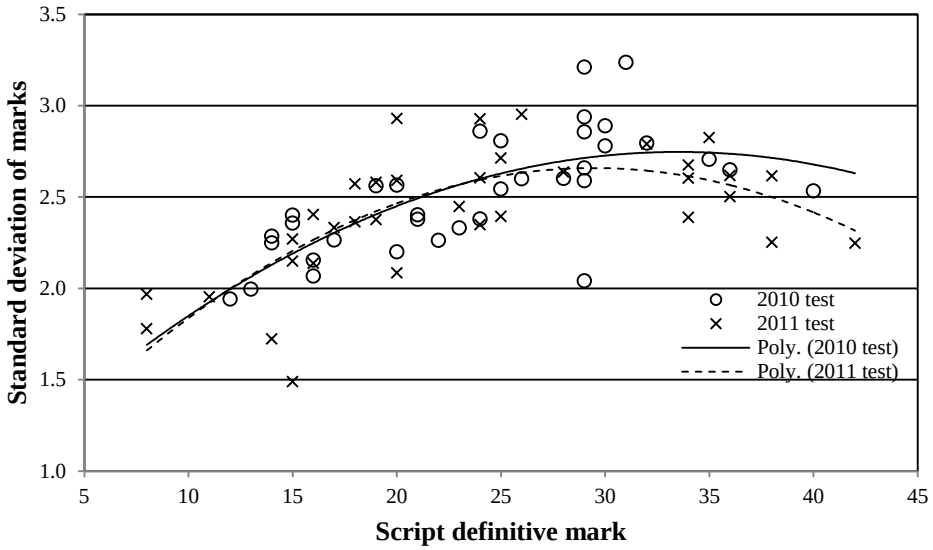

Figure 5 Variation of the standard deviation of marks from markers assigned to the standardisation and benchmarking scripts with script definitive mark

The average standard deviation of the difference marks between markers' marks and script definitive marks is estimated to be 2.88 for the 2010 test and 3.17 for the 2011 test, representing about 7 per cent of the maximum available mark (43) on the tests, 25 per cent of the level bandwidth (defined as the interval between two adjacent level boundary scores, which is about 12 marks), and 12 per cent of the mean mark on the tests (excluding the spelling test section; see Table 1).

The following second-order polynomial function was fitted to the distribution of standard deviations of marks for all the standardisation and benchmarking scripts for each of the tests:

$$y = \begin{cases} -0.0016x^2 + 0.1087x + 0.952 \quad (R^2 = 0.52, \text{ 2010 test}), \\ \\ -0.0022x^2 + 0.128x + 0.7764 \quad (R^2 = 0.55, \text{ 2011 test}), \end{cases} \tag{1}$$

where y is the model fitted standard deviation, and *x* is the script definitive mark. The model fitted standard deviations will provide a basis for investigating variability of marking over time (see discussion below).

3.2 Variability of marking over time

If the actual standard deviations of the marks for the scripts in Figure 5 are compared with the model fitted values, it is possible to investigate variability in marking between the markers over time by introducing a relative measure of mark variation for a script

– the scaled standard deviation of marks SD_{scaled}, which is defined as the ratio of the standard deviation of marks SD on a script to the model fitted value SD_{m}:

$$SD_{\text{scaled}} = \frac{SD}{SD_{\text{m}}}. \quad (2)$$

Figure 6 shows how the scaled standard deviations of markers' marks assigned to the standardisation and benchmarking scripts vary with script definitive marks. The scaled standard deviation takes into account the effect of script definitive mark on the standard deviation of marks. As is clear from Figure 6, the scaled standard deviations remain relatively constant for scripts with different definitive marks.

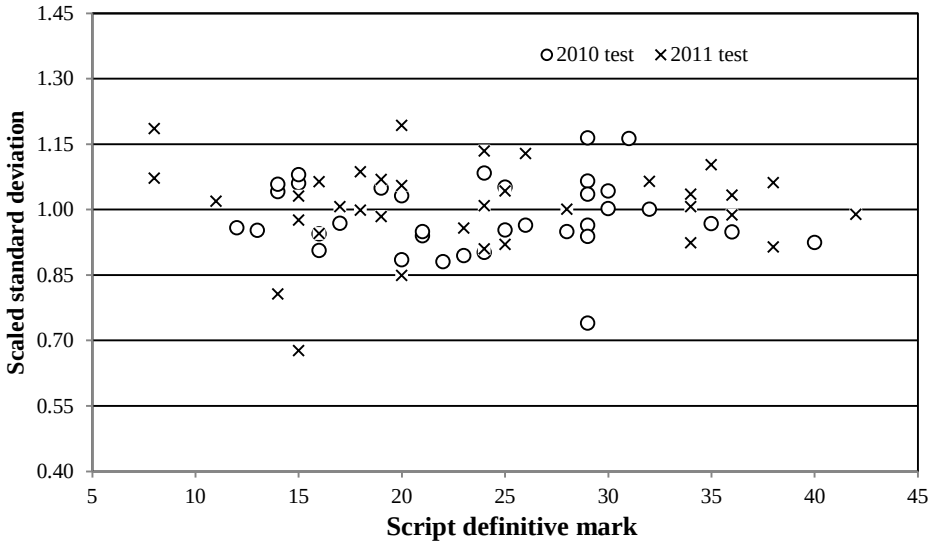

Figure 6 Variation of the scaled standard deviation of marks from markers assigned to the standardisation and benchmarking scripts with script definitive mark

Figure 7 illustrates the average scaled standard deviations of marks from markers on the standardisation and benchmarking scripts. A substantial decrease in the average scaled standard deviations over time (i.e. from Standardisation 1 to Benchmarking 1, from Benchmarking 1 to Benchmarking 2) would suggest a decrease in variability of marking between the markers over time. It appears from Figure 7 that variability in marking at the overall script level was relatively stable for the markers who marked the 2010 and 2011 tests.

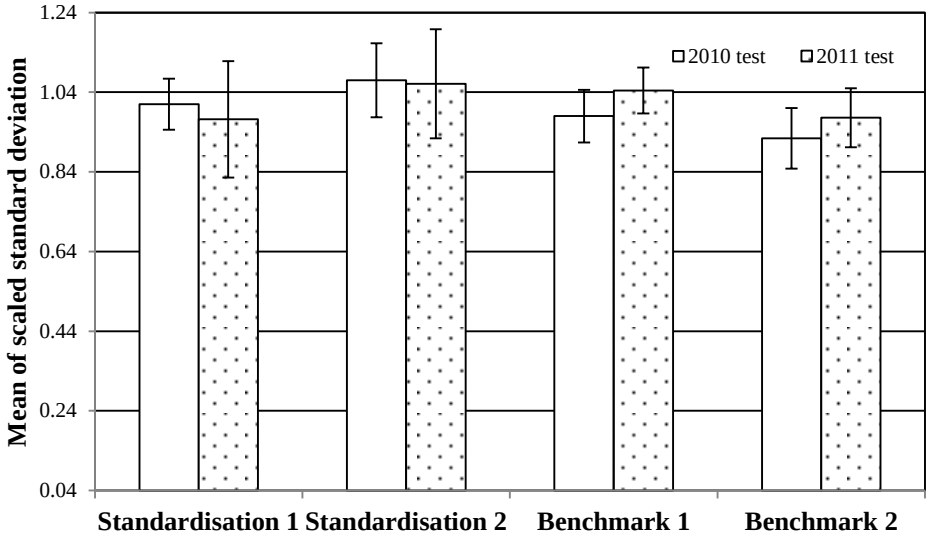

Figure 7 Distribution of the average scaled standard deviation of marks from markers assigned to the standardisation and benchmarking scripts

3.3 Variability of marking at marking strand level

Variability in marking between markers at task level and marking strand level were investigated through comparison between markers' marks on individual marking strands and the strand definitive marks. Figure 8 depicts how the difference marks between markers' mean mark on individual marking strand vary with the strand definitive mark. As can be seen, for large tariff marking strands such as the composition and effect (CE) strand associated with both the short and the long tasks, markers tend to 'over-mark' lower quality work and 'under-mark' higher quality work. The patterns of variability in the difference marks are similar for both the 2010 and 2011 tests.

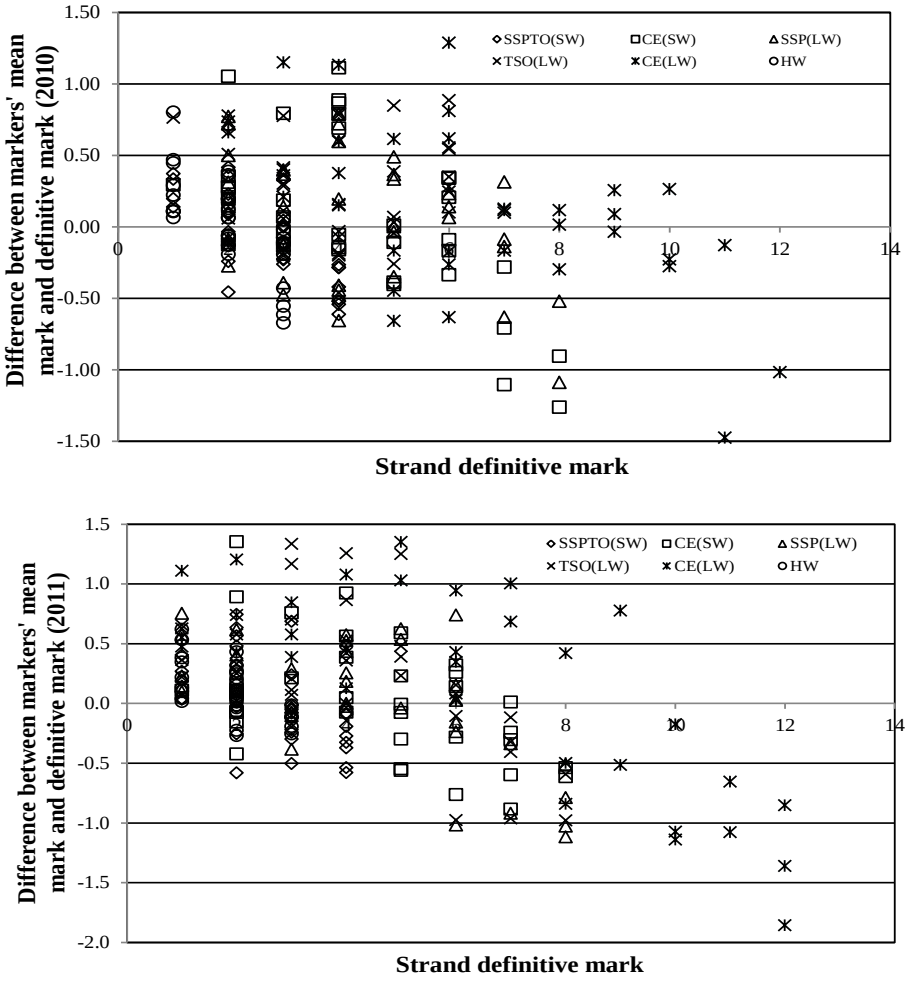

Figure 8 Variation of difference marks between markers' mean marks on individual marking strands and strands' definitive marks with strand definitive mark

Figure 9 shows how the standard deviations of markers' marks on individual marking strands vary with strand definitive marks. There is substantial variability in marking between markers at marking strand level. For marking strands with similar definitive marks, variation in marks is generally larger for higher tariff marking strands than for lower tariff marking strands. Similar to the distribution of the standard deviations of marks at script level, for higher tariff marking strands, the standard deviation of marks generally increases for those scripts with marks towards the lower end of the strand definitive mark, plateaus in the middle and then decreases towards the upper end of the strand definitive mark. The composition and effect marking strand for the long task has a maximum of 12 marks, for which the pattern of variability in the marks is clearer than that for other marking strands. It should be noted that for handwriting, the maximum mark is only 3 marks and therefore small variability in marking would be expected. It is apparent that the patterns of variability in marking are similar for the markers who marked both the 2010 and 2011 tests.

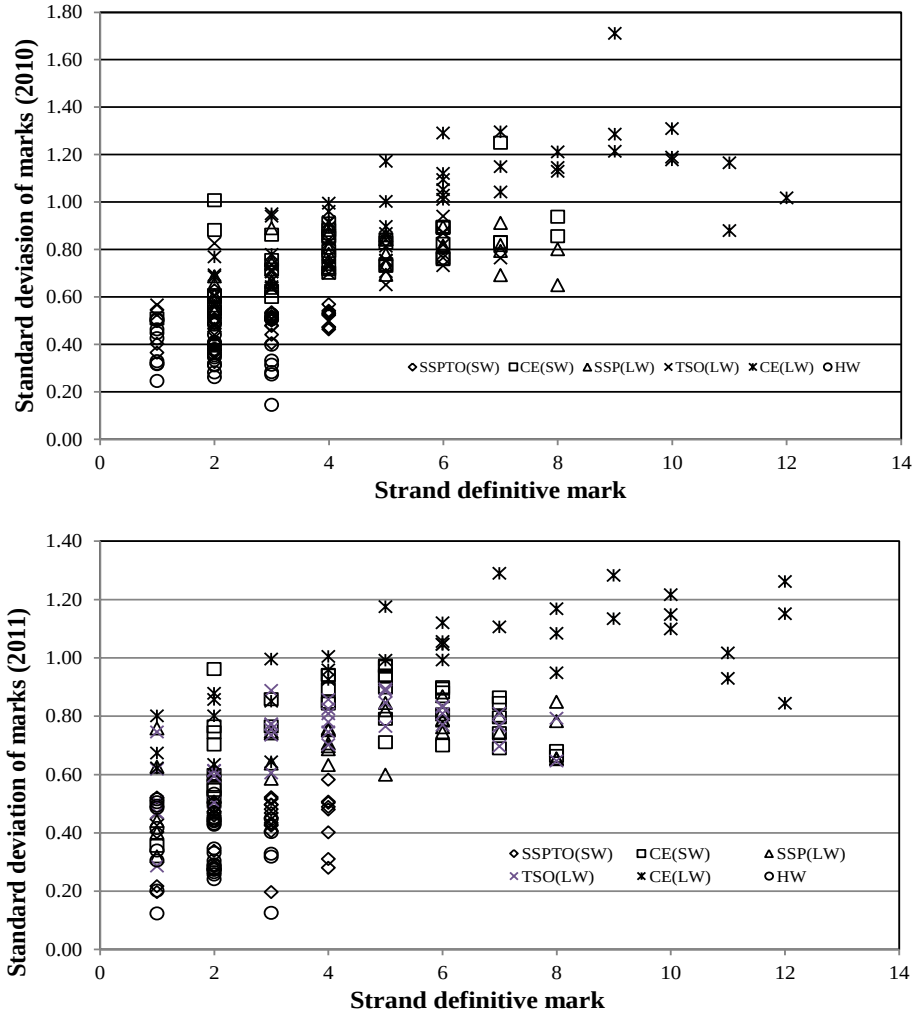

Figure 9 Variation of the standard deviation of marks assigned to individual marking strands by markers with strand definitive marks

3.4 Stability of consistency in marking over time for individual markers

To study the stability of consistency in marking over time for individual markers, the absolute mark difference AMD for a marker, which is defined as the sum of the absolute differences between the marker's marks given to individual scripts and the scripts' definitive marks for the standardisation scripts or the benchmarking scripts, is used as a measure of inconsistency in their marking:

$$AMD = \sum_{i=1}^{N} \left| m_i - m_{\text{definitive},i} \right|, \quad (3)$$

where m_i is the marker's mark on script *i*, $m_{\text{definitive},i}$ is the script definitive mark, and *N* is the number of standardisation or benchmarking scripts. Large AMD would indicate

lower level of agreement between the marker's marks on the scripts and the scripts' definitive marks, while small AMD would suggest a higher level of agreement between the marker's marks and scripts' definitive marks. High levels of correlations between the markers' AMDs would suggest high stability of consistency in marking over time for individual markers. Tables 3 and 4 list the correlations between AMDs for the markers who marked the 2010 and 2011 Key Stage 2 English writing tests. The AMDs are correlated, although the strength of the correlation is small to medium. This suggests that there was a proportion of markers who were consistent in their marking over time –marking consistently either accurately or inaccurately.

Table 3 Correlations between markers' AMDs for the 2010 Key Stage 2 English writing test

	Standardisation 1	Benchmarking 1	Benchmarking 2
Standardisation 1	1	0.338**	0.272**
Benchmarking 1		1	0.364**
Benchmarking 2			1

** significant at $p < 0.01$

Table 4 Correlations between markers' AMDs for the 2011 Key Stage 2 English writing test

	Standardisation 1	Benchmarking 1	Benchmarking 2
Standardisation 1	1	0.346**	0.342**
Benchmarking 1		1	0.372**
Benchmarking 2			1

** significant at $p < 0.01$

3.5 Marker-related and test-related standard error of measurement in test scores

Test internal consistency reliability

Figure 10 shows the mark distributions for the samples of pupils that were used to generate distributions of marks on individual marking strands as well as the overall test for the 2010 and 2011 tests. The means and standard deviations of the distributions are 25.39 and 6.82 for the 2010 test, and 24.73 and 6.81 for the 2011

test. The marks are normally distributed with a good dispersion, suggesting that the tests were targeting the pupils well and the available marks were fully used.

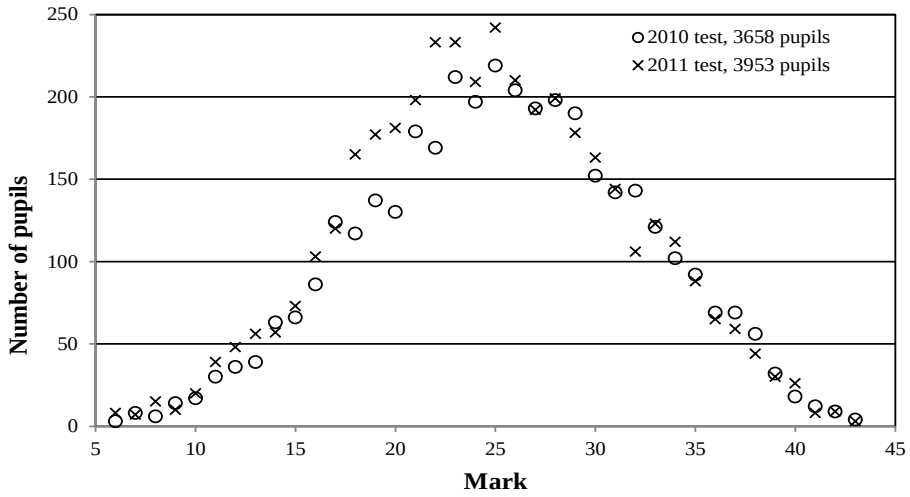

Figure 10 Mark distributions for pupils from the 2010 and 2011 Key Stage 2 English writing tests for whom detailed marks on individual marking strands were collected

Table 5 lists the sample size, average marks and standard deviations of the mark distributions for individual marking strands and the overall tests. As indicated earlier, these mark distributions and the strand level mark data were used to assist standards setting and to evaluate the performance of the tasks and the overall test. The two tests performed similarly for the two samples at both individual marking strand level and the overall test level.

Table 5 Number of pupils for whom detailed marks on individual marking strands were recorded from the 2010 and 2011 Key Stage 2 English writing tests and strand and test statistics

		2010		2011	
	Maximum mark	Mean	SD	Mean	SD
SSPTO (S)	4	2.81	0.75	2.79	0.75
CE (S)	8	4.68	1.41	4.57	1.53
SPP (L)	8	4.59	1.48	4.50	1.41
TSO (L)	8	4.49	1.37	4.34	1.37
CE (L)	12	6.76	2.21	6.49	2.16
HW (L)	3	2.06	0.64	2.03	0.61
Number of pupils		3,658		3,953	
Test mean score		25.39		24.73	
Test SD		6.82		6.81	
Cronbach's alpha		0.893		0.895	

Traditionally, items in any test consist of separate questions requiring separate responses from the test-takers, although the questions may share the same stimulus. For the Key Stage 2 English writing tests, if individual marking strands specified are treated as separate items associated with either the long task or the short task, then an internal consistency reliability measure similar to Cronbach's alpha can be estimated. Cronbach's alpha was used here as an internal consistency reliability measure, and was estimated to be about 0.893 for the 2010 test and 0.895 for the 2011 test, which are almost the same. Because of inconsistency in marking between the markers, there is an error component in the strand scores, and this will affect the value of Cronbach's alpha, which is estimated based on inter-strand correlations. To investigate the effect on Cronbach's alpha of the variability of marking between markers, a simulation was undertaken. In the simulation, a random error score, with a normal distribution and standard deviation assumed to be that of the corresponding strand difference scores on the standardisation and benchmarking scripts, was added to the individual strand scores, and Cronbach's alpha was re-estimated. The results from the simulation indicated that Cronbach's alpha was underestimated by about 5 per cent. Therefore, had there been no error from unreliability of marking, Cronbach's alpha would be well over 0.90 for the two tests.

Standard error of measurement in test scores

Various measures have been used to quantify unreliability in marking between markers (see Bramley, 2007; Bramley and Dhawan, 2012). These include, among others, agreement rate between markers' marks or between markers' marks and definitive marks, correlations between markers' marks or between markers' marks and definitive marks, and intraclass correlations between markers' marks. The standard error of measurement (SEM), which is defined as the average of the standard deviations of the error scores for individual test-takers, is also widely used as a measure of unreliability in test scores (Harvill, 1991; Wiliam, 2001; Bachman, 2004). The SEM can be used to calculate confidence intervals for observed scores or true scores (Harvill, 1991), which can be interpreted more easily than most of the other reliability measures. Table 6 shows the standard error of measurement in test scores estimated for the two tests, based on inconsistency in marking between markers and the internal consistency reliability of the tests.

Table 6 Test-related and marker-related standard error of measurement in test scores for the Key Stage 2 English writing tests administered in 2010 and 2011

	Test-related SEM (SEM_{test})	Marker-related SEM (SEM_{marker})	SEM_{test}/SEM_{marker}
2010	2.23	2.50	0.89
2011	2.21	2.42	0.91

The marker-related SEM (SEM_{marker}) in test scores was initially assumed to be the standard deviation of the difference marks between the markers' marks on the standardisation and benchmark scripts, and the script definitive marks (2.88 for the 2010 test, and 3.09 for the 2011 test). However, as discussed above, the difference between markers' mean marks on a script and the script definitive mark for some standardisation and benchmarking scripts can be considerable and because it is the inconsistency between the markers that affect the reliability of the final test scores, a more appropriate measure of marker-related SEM would be the standard deviation of the difference scores between markers' marks and their mean marks ($SD_{difference,mean}$) on the scripts:

$$SEM_{\text{marker}} = SD_{\text{difference, mean}} . \quad (4)$$

Based on the distribution of difference marks between markers' marks and markers' mean marks, the marker-related SEM was estimated to be 2.50 for the 2010 test and 2.42 for the 2011 test, which is about 20 per cent of the level bandwidth. These

estimates were considerably smaller than estimates based on the difference marks between the markers' marks and scripts' definitive marks.

The test-related SEM (SEM_{test}) was estimated from the value of the internal consistency reliability *r* (Cronbach's alpha) and the standard division of the mark distribution SD_{test} shown in Table 5:

$$SEM_{\text{test}} = SD_{\text{test}}\sqrt{1-r}\,. \tag{5}$$

The test-related SEM in test scores was estimated to be 2.23 for the 2010 test and 2.21 for the 2011 test (see Table 6). It is clear that the marker-related SEM is larger than the test-related SEM for both tests, with the ratio SEM_{test}/SEM_{marker} being 0.89 for the 2010 test and 0.91 for the 2011 test. This would suggest that pupils were more likely to get a different result if their work was marked by a different marker than if they took a different, parallel writing test. The overall SEM in test scores that incorporates both marker-related unreliability and test-related unreliability will be higher than the larger of the two SEMs.

4 Conclusions and Discussion

Findings from the present study indicated that there was considerable variability in the marking of the scripts used for standardisation and benchmarking for the Key Stage 2 National Curriculum English writing tests administered in 2010 and 2011 at both question paper level and individual task/marking strand level, which reflects the nature of the tests. These findings are broadly consistent with the findings from other marking reliability studies of assessments that require extended responses from the test-takers (see Bramley and Dhawan, 2012; Dhawan and Bramley, 2012). It is realised that the sample size of the scripts used for standardisation and benchmarking in this study is small, and caution has therefore to be exercised when drawing any conclusions from the findings about the reliability of marking of the Key Stage 2 English writing test. However, given that these scripts cover a wide range of definitive marks and hence pupils' work of varying quality, the variability of the difference marks between the markers' marks on the standardisation and benchmarking scripts, and the script definitive marks, or the difference marks between the markers' marks and the markers' mean marks on the scripts, would provide a reasonably good measure of unreliability attributable to marking inconsistency in the final test scores. The present study has provided a better understanding of the level of marking reliability of the Key Stage 2 English writing tests.

It was shown that the quality of pupils' work to an extent affected the consistency in marking between markers at both question paper level and individual task or marking strand level, with low and high performing work generally more consistently marked than medium performing work. This might be expected as it might be easier for markers to reach consensus regarding the quality of pupils' work and the application of the marking criteria for both poor and high quality work. For work of medium quality, the 'zone of uncertainty' could be large, and there is potential for different markers to award a different mark to the same piece of work (see Sweiry, 2012). It is, however, to be noted that for pupils' work of similar quality (as reflected by the similar script definitive marks), there can be substantial variation in the variability of marking between markers for different scripts. This variation in marking for scripts of similar quality or definitive marks may partly reflect the variation associated with the quality of specific aspects of the writing skills being assessed.

Results from this study indicated that markers' mean marks on scripts were generally higher than the scripts' definitive marks for scripts with lower definitive marks, but lower for scripts with higher definitive marks. Conventionally, the definitive marks determined by a small panel of senior markers for scripts used for marker standardisation and marking quality monitoring are treated as 'gold standards', and the difference between markers' marks and scripts' definitive marks are used to assess the quality of marking for individual markers. The considerable difference

between markers' mean mark on a script and the script definitive mark for some of the standardisation and benchmarking scripts needs to be taken into account when estimating the inter-marker reliability of the tests, since the number of normal markers involved in live marking is generally substantially larger than the number of senior markers involved in determining the definitive marks, and it is the inter-marker variability in marking between the normal markers that affects the reliability of the final test scores for the test-takers. It has been shown that the marker-related SEM in test scores estimated from the distribution of the difference marks between markers' marks and markers' mean marks on the standardisation and benchmarking scripts is considerably smaller than that estimated from the distribution of the difference marks between the markers' marks and the scripts' definitive marks. The difference between markers' mean marks and script definitive marks would also point to the possibility of exploring alternative definitions of 'definitive' mark for scripts used for monitoring marking quality. One possibility would be to use the mean marks from the markers on the standardisation and benchmarking scripts as 'definitive marks' when assessing the quality of marking for individual markers. This approach would be consistent with the conception of 'true scores' used in classical test theory. Dhawan and Bramley (2012) discuss other possible alternative definitions of 'definitive' mark for a script. It should also be noted that the markers knew when they were marking the standardisation and benchmarking scripts, and this might have an effect on their marking behaviour when marking these scripts, and hence affect their mean marks.

It has also been shown in the present study that there was no substantial variation in marking inconsistency over time for the markers marking the 2010 and 2011 Key Stage 2 English writing tests. It was found that individual markers to a degree tended to be stable in their marking over time.

It has been demonstrated that marker-related standard error of measurement for the 2010 and 2011 Key Stage 2 English writing tests was higher than test-related standard error of measurement in test scores, suggesting that inconsistency in marking between markers would contribute a larger part to the unreliability of the results from the test than the unreliability associated with test tasks. This reflects the nature of this type of test: the responses from the test-takers are diverse, and the level of marking demand placed on markers is high when they make subjective judgement of the quality of the test-takers' work and apply the marking criteria. Substantial research has been undertaken to study the many factors that can affect marking demand (Bramley, 2008; Suto et al., 2008; Ahmed and Pollitt, 2011; Black et al., 2011; Sweiry, 2012). These factors include, among others, the nature of the assessment tasks, the nature of pupils' responses, the expertise of the markers, and the nature of the mark scheme. These factors will result in inconsistency in the perceived importance of the specific knowledge, skills and understanding in reflecting the learning requirements by the markers, the interpretation of the mark scheme and

the quality of pupils' work and the application of the marking criteria between the markers.

Although the level of reliability that can be practically achieved for an assessment to a large extent is affected by the nature of the assessment and its validity requirement, there has been substantial research to explore ways of improving the reliability of marking (Brooks, 2004; Meadows and Billington, 2005; Ahmed and Pollitt, 2011; Crisp, 2010; Kim and Wilson, 2009; Dhawan and Bramley, 2012; Blood, 2011; Pollitt, 2012; Tisi et al., 2013). These include the selection of effective markers, the development of effective mark schemes, the training of markers, procedures used for marking pupils' work, and the monitoring of marking quality during marking.

The present study used the distributions of the difference marks between markers' marks on the standardisation and benchmarking scripts and the script definitive marks (or markers' script mean marks) and Cronbach's alpha to investigate marker-related and test-related unreliability in test scores separately. As a result, it was not possible to estimate the overall measurement error in the test scores. Further work involving the use of other techniques such as Rasch modelling and generalizability theory (G-theory) to analyse the data would be useful (Johnson and Johnson, 2012; Baird et al., 2012). One of the advantages of G-theory analysis is that it allows the estimation of the overall measurement error in test scores as well as the relative contributions from the different sources of unreliability to the overall measurement error. Once the overall measurement error in test scores is estimated, it will be possible to estimate the percentage of pupils who were awarded an inaccurate attainment level from the level boundary scores, i.e. the misclassification rate (see He et al., 2012b).

It is hoped that findings from the present study will be useful for the development and investigation of marking reliability of English writing tests and other assessments that require extended responses from the test-takers in general.

References

Ahmed, A. and Pollitt, A. (2011) Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy and Practice*, **18** (3), 259–78.

Bachman, L. (2004) *Statistical Analyses for Language Assessment*. Cambridge, Cambridge University Press.

Baird, J., Greatorex, J. and Bell, J. (2004) What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy and Practice*, **11**, 331–48.

Baird, J., Hayes, M., Johnson, R., Johnson, S. and Lamprianou, I. (2012) Marker effect and examination reliability: a comparative exploration from the perspectives of generalizability theory, Rasch modelling and multi-level modelling. Available online at: http://www.ofqual.gov.uk/files/2013-01-21-marker-effects-and-examination-reliability.pdf.

Baker, E., Ayers, P., O'Neill, H., Choi, K., Sawyer, W., Sylvester, R. and Carroll, B. (2008) KS3 English test marker study in Australia. Report to the National Assessment Agency of England. London, Qualifications and Curriculum Authority.

Benton, T. (2006) Exploring the importance of graders in determining pupils' examination results using cross-classified multilevel modelling. Paper presented at the European Conference on Educational Research, 14th September, University of Geneva.

Bew, P. (2011) *Independent Review of Key Stage 2 Testing, Assessment and Accountability: Final Report*. Available at: https://www.education.gov.uk/publications/eOrderingDownload/Review-KS2-Testing_final-report.pdf (accessed 2nd March 2013).

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

Blood, I. (2011) Automated essay scoring: a literature review. *Working Papers in TESOL & Applied Linguistics*, **11**, 40–64. Available at: http://journals.tc-library.org/index.php/tesol/article/download/745/470 (accessed 2nd March 2013).

Bramley, T. (2007) Quantifying marker agreement: terminology, statistics and issues. *Research Matters*, **4**, 22–8.

Bramley, T. (2008) Mark scheme features associated with different levels of marker agreement. Paper presented at the annual conference of the British Educational Research Association, Edinburgh. Available at: www.cambridgeassessment.org.uk/ca/digitalAssets/171183_TB_MSfeatures_BERA08.pdf (accessed 2nd March 2013).

Bramley, T. and Dhawan, V. (2012) Estimates of reliability of qualifications. In Opposs, D. and He, Q. (eds) *Ofqual's Reliability Compendium*, pp. 217–320. Coventry, Office of Qualifications and Examinations Regulation.

Brennan, R. (2001) An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, **38**, 295–317.

Brooks, V. (2004) Double marking revisited. *British Journal of Educational Studies*, **52**, 29–46.

Crisp, V. (2010) Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education*, **36**, 1–21.

Department for Education (DfE) (2011) *The Framework for the National Curriculum: A Report by the Expert Panel for the National Curriculum Review*. Available at: www.education.gov.uk/publications/eOrderingDownload/NCR-Expert%20Panel%20Report.pdf (accessed 2nd March 2013).

Dhawan, V. and Bramley, T. (2012) Estimation of inter-rater reliability. Available online at: http://www.ofqual.gov.uk/files/2013-01-17-ca-estimation-of-inter-rater-reliability-report.pdf.

Fowles, D. (2009) How reliable is marking in GCSE English? *English in Education*, **43**, 50–67.

Haertel, E. (2006) Reliability. In Brennan, R. (ed.) *Educational Measurement* (4th edn), pp. 65–110. Westport, CT: American Council on Education/Praeger.

Harvill, L. (1991) An NCME instructional module on standard error of measurement. *Educational Measurement: Issues and Practice*, **10**, 33–41.

He, Q., Anwyll, S., Glanville, M. and Opposs, D. (2012a) An investigation of measurement invariance of the Key Stage 2 National Curriculum science sampling test in England. *Research Papers in Education*, DOI:10.1080/02671522.2012.74213.

He, Q., Hayes, M. and Wiliam, D. (2012b) Classification accuracy in results from KS2 National Curriculum tests. In Opposs, D. and He, Q. (eds) *Ofqual's Reliability Compendium*, pp. 91–105. Coventry, Office of Qualifications and Examinations Regulation.

He, Q. and Opposs, D. (2012) The reliability of results from national tests, public examinations, and vocational qualifications in England. *Educational Research and Evaluation*, **18**, 779–99.

Headmasters' and Headmistresses' Conference (2012) *England's 'Examinations Industry': Deterioration and Decay*. Available at: www.hmc.org.uk/wp-content/uploads/2012/09/HMC-Report-on-English-Exams-9-12-v-13.pdf (accessed 2nd March 2013).

House of Commons Children, Schools and Families Committee (2008) *Testing and Assessment, Third Report*. Available at: www.publications.parliament.uk/pa/cm200708/cmselect/cmchilsch/169/169.pdf (accessed 2nd March 2013).

Isaacs, T. (2010) Educational assessment in England. *Assessment in Education: Principles, Policy & Practice*, **17**, 315–34.

Johnson, M., Nadas, R. and Bell, J. (2010) Marking essays on screen: an investigation into the reliability of marking extended subjective texts. *British Journal of Education Technology*, **41**, 814–26.

Johnson, S. and Johnson, R. (2012). Component reliability in GCSE and GCE. In Opposs, D. and He, Q. (eds) *Ofqual's Reliability Compendium*, pp. 141–216. Coventry, Office of Qualifications and Examinations Regulation.

Kim, S. and Wilson, M. (2009) A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement*, **10**, 408–23.

Meadows, M. and Billington, L. (2005) *A Review of the Literature on Marking Reliability*. Manchester, AQA.

Murphy, R. (1982) A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, **52**, 58–63.

Murray, J. (2011) Headteachers reject 'appalling' marking of this year's Sats tests. *The Guardian*, 11th July. Available at: www.guardian.co.uk/education/2011/jul/11/sats-marking-rejected-by-headteachers (accessed 28th February 2013).

Newton, P. (2009) The reliability of results from national curriculum testing in England. *Educational Research*, **51**, 181–212.

Office of Qualifications and Examinations Regulation (Ofqual) (2012) *GCSE English 2012*. Available at: www.ofqual.gov.uk/files/2012-11-02-gcse-english-final-report-and-appendices.pdf (accessed 2nd March 2013).

Opposs, D. and He, Q. (eds) (2012) *Ofqual's Reliability Compendium*. Coventry, Office of Qualifications and Examinations Regulation.

Pollitt, A. (2012) The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice*, **19**, 281–300.

Qualifications and Curriculum Authority (QCA) (1999) *The National Curriculum for England, Key Stages 1–4*. London, Department for Education and Employment and Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority (QCA) (2009) Test development, level setting and maintaining standards. Available at: http://webarchive.nationalarchives.gov.uk/20090608182316/testsandexams.qca.org.uk/18939.aspx (accessed 2nd March 2013).

Qualifications and Curriculum Development Agency (QCDA) (2011) *English Tests: Mark Schemes – Reading, Writing and Spelling Tests*. Available at: www.satstestsonline.co.uk/past_papers/2011_english_mark_scheme.pdf (accessed 2nd March 2013).

Royal-Dawson, L. (2005) Is teaching experience a necessary condition for markers of Key Stage 3 English? London, Qualifications and Curriculum Authority.

Royal-Dawson, L. and Baird, J. (2009) The impact of teaching experience upon marking reliability in Key Stage 3 English. *Educational Measurement: Issues and Practice*, **28**, 2–8.

Royal-Dawson, L., Leckie, G. and Baird, J. (2008) Marking reliability of the 2008 National Curriculum tests. Available at: http://archive.teachfind.com/qcda/orderline.qcda.gov.uk/gempdf/184962531X/QCDA104981_marking_reliability_tests_2008.pdf (accessed 2nd March 2013).

Suto, I., Crisp, V. and Greatorex, J. (2008) Investigating the judgemental marking process: an overview of our recent research. *Research Matters*, **5**, 6–9.

Suto, I., Nadas, R. and Bell, J. (2011) Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, **26**, 21–51.

Sweiry, E. (2012) Conceptualising and minimising marking demand in selected and constructed response questions. Paper presented at the Association for Educational Assessment – Europe (AEA-Europe), Berlin, Germany.

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

University of Cambridge Local Examinations Syndicate (2002) QCA KS3 English 2003: Pretest, June 2002. Marker reliability study. London, Qualifications and Curriculum Authority.

Ward, H. (2009) Primaries still concerned over marking of KS2 writing tests. *Times Educational Supplement*, 17th July. Available at: www.tes.co.uk/article.aspx?storycode=6017939 (accessed 28th February 2013).

Whetton, C. (2009) A brief history of a testing time: National Curriculum assessment in England 1989–2008. *Educational Research*, 51, 137–59.

Wiliam, D. (2001) Reliability, validity, and all that jazz. *Education*, **29**, 17–21.

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 2014

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