QUALITY OF MARKING OF KS3 ENGLISH

Lucy Royal-Dawson

1. INTRODUCTION

In 2004, Key Stage 3 English was component marked for the first time. Reading markers marked the two reading papers and writing markers marked the two writing papers. Prior to this, all four papers were marked by the same marker. There was a perceived expectation that component marking might increase marking reliability. To judge whether this was the case, three investigations of marking in 2003 and 2004 were undertaken. First, standardisation and first sample marks from three Senior Markers' teams, one from 2003 and two from 2004, were explored to compare the mark differences between the supervising marker and the markers. Second, requests for individual reviews of pupils' marks were analysed with regard to the proportion of mark changes that resulted from reviews. Third, the number of group reviews that resulted in a level change this year compared with last year was estimated from a sample of scripts.

2. STANDARDISATION AND FIRST SAMPLE MARKS

Background

Markers are trained to mark through standardisation. Immediately after training, they send scripts they have marked, the standardisation sample, to their team leader who checks the accuracy of the marking. Marker acceptability is judged by reference to the absolute mark difference (AMD) between the team leader's and marker's mark allocated to a pupil. The AMDs are calculated at strand level[1] for the two writing papers (writing and Shakespeare writing) for each pupil and at total mark level for each of the two reading papers (reading and Shakespeare reading) for each pupil. The pupil AMDs are summed across all pupils in the sample and marking acceptability is graded according to the size of the sample AMD. Once a marker is deemed acceptable, he or she can start marking. If a marker does not meet the acceptability criterion, the team leader requests further samples for checking. The quality of marking is monitored in the same way during the marking period by the team leaders through two further samples of scripts from each marker, the first early on in the marking period and the final at about three-quarters of the way through the allocation.

Method

The current study used the standardisation and first sample forms from three Senior Markers' teams, one from 2003 and two from 2004 representing reading and writing. The teams were selected because their paperwork was readily available at short notice. Only markers who had forms for both the standardisation and first samples were used to provide consistency between standardisation and first sample. The size and composition of the teams used in the study are given in Table 1 below. The paperwork for the teams in 2003 was not complete in that not every marker had forms for both reading and writing scripts.

Table 1 Size and composition of the marking teams in the study

	2003		2004	
	Reading	Writing	Reading	Writing
Number of team leaders	8	8	6	6
Number of markers	53	57	49	38

[1] The writing strands are: sentence structure and punctuation (SSP), text structure and organisation (TSO) and composition and effect (C&E); the Shakespeare writing strands are: sentence structure, punctuation and text organisation (SSPTO), composition and effect (C&E) and spelling.

To compare the quality of marking between the two years, the following measures were calculated from the mark differences between the team leaders and markers recorded for both the standardisation and first samples:

- The proportion of agreements in marks between a team leader and a marker.
- The proportion of agreements within one mark (that is, the same mark or one mark either side) between a team leader and a marker.
- The mean pupil AMD and its standard deviation across all markers for each of the four papers.

The agreement rate provides an indication of the consistency of marking between team leaders and markers. The mean pupil AMD indicates the average size of the discrepancy in marking between all markers and their team leaders. From standardisation to the first sample, markers are expected to improve, which would be indicated by a drop in the mean pupil AMD. The above measures were compared in two ways: between the two years of interest to investigate differences in marking quality, and between standardisation and first sample to investigate whether standardisation had led to improvement in marking in both years.

Results: comparison between years

The comparison between the two years of interest indicated that, in the case of the standardisation sample, the proportion of agreements in marks between the team leaders and markers in 2004 was higher than in 2003 for both reading and writing. Referring to Table 1a in the Appendix as an example of the comparisons, the proportion of agreements in marks on the reading paper in 2003 was 12.7%, but it rose to 16.1% in 2004. Similar increases were observed for the Shakespeare reading paper (also Table 1a) and the two writing papers (Table 2a). Accordingly, the mean pupil AMDs for both components were slightly lower in 2004 than in 2003 (see Tables 1b and 2b in the Appendix). For example, the mean pupil AMD for the writing paper in 2003 was 4.06 marks, but in 2003 it dropped to 3.17 marks in 2004 (Table 2b). The decreases in the mean pupil AMD observed between the two years were less than one mark, which is of little educational significance.

The differences between the two years were more varied for the first sample. The proportion of agreements in marks in the reading component was lower in 2004 than in 2003 (Table 3a). In the writing component, some strands had a higher proportion of agreement in 2003 and others in 2004 (Table 4a). The mean pupil AMD was slightly lower in 2004 in the reading paper, but not in the Shakespeare reading paper (Table 3b). Both papers of the writing component indicated higher mean pupil AMD in 2004 (Table 4b). The differences in mean were all less than 0.2 of a mark, again suggesting little educational significance.

Results: comparison of standardisation and first samples

The comparison between the measures from standardisation to first sample for both years indicated the expected improvement in marking. Referring to Tables 1 and 3 in the Appendix which relate to the reading papers, increases in the proportion of agreements and decreases in the mean pupil AMD were notable in both years. The writing papers show the same pattern of improvement in Tables 2 and 4 in the Appendix.

Conclusion

This investigation of the standardisation and first samples from three opportunistically chosen teams of markers indicated few differences between the two years suggesting the quality of

marking did not change greatly, and the process of standardisation was efficient. Whilst the change to component marking might have been expected to lead to a greater increase in the quality of marking, these results suggest there was little improvement to be made on whole subject marking.

3. INDIVIDUAL REVIEWS

Background

Schools request marking reviews for individual pupils (known as R2s) if they suspect the marking has resulted in a pupil being assigned the wrong level. The request must be supported with reference to the question or strand (see footnote 1) believed to be incorrectly marked. The External Marking Agency (EMA) assigns R2s to specially trained re-markers.

The EMA's procedure for R2s in 2003 and 2004 only differed to take account of the separation and later re-combination of reading and writing papers. All other aspects of the review process from training to reporting remained the same. It is important to note, however, that some schools may have approached R2s differently in 2004. In previous years, schools knew their pupils' levels at the time of making a request, and were able to specify where they felt the marking was errant, thus making well targeted requests. Indeed, a study in 2003 indicated that approximately two-thirds of the requests were upheld and resulted in a pupil receiving a mark change (Schermbrucker, 2004). In 2004, however, some schools did not know their pupils' correct levels because of inaccuracies in the QCA results database at the time of making R2 requests. These schools were presented with a contradiction between the marked paper they were holding and the incorrect level on the database. Some schools would have made individual review requests based on this incorrect information. This suggests they used R2s as a vehicle for addressing their confusion resulting in less well targeted queries about marking.

Method

The original marks and re-marks from samples of R2 scripts in 2004 were keyed. This allowed the proportion of mark changes to be calculated for comparison with figures from 2003.

Results

The number of separate schools making a request for an R2 in 2003 was 611, and in 2004 it was 658, an increase of 7.7%. This 2004 figure was calculated by summing schools that made a reading only, writing only or a reading and writing request, as shown in bold in Table 2. Because the number of pupils involved in 2003 was not disaggregated into the number of reading only, writing only or reading and writing requests, it is difficult to compare the pupil numbers between the two years.

Table 2 Number of requests for individual reviews made by schools and the number of pupils involved in 2003 and 2004

	2003	**2004**		
	Reading or writing or both	Reading (reading only)	Writing (writing only)	Both reading and writing
Schools	611	542 (**84**)	574 (**116**)	**458**
Pupils	5,331	5,090	5,542	*

* This figure is not easily available because records have to be matched at pupil name level: there is no ready pupil identifier.

The proportion of R2 requests that resulted in a mark change dropped in all four papers in 2004, shown in Table 3. The reading paper in particular saw a dramatic decrease in mark changes.

Table 3 The proportion of R2s that resulted in a mark change in 2003 and 2004

	2003		2004	
Paper	No. in sample	% of mark changes	No. in sample	% of mark changes
Reading	2,744	72.3	674	35.3
Shakespeare reading	990	68.2	647	56.4
Writing	3,090	75.4	1,373	63.8
Shakespeare writing	1,774	70.8	1,039	54.0

Conclusion

The drop in the proportion of R2s resulting in a mark change suggested the original marking in 2004 was defensible more often than it was in 2003. In the past, the proportion has been higher probably because schools were able to make well targeted requests, but in 2004, some schools made requests based on incorrect information, which yielded requests which were less well targeted. These latter requests would not have been defensible as often as the well targeted requests.

4. GROUP REVIEWS

Background

Schools request a review of marking for all pupils if they suspect the marking is erratic or consistently too severe or too lenient. The school must identify a sample of pupils whose scripts typify the perceived marking errors when the whole cohort is sent for review, known as a Group Review (GR). The EMA first assigns GRs to specially trained reviewers who review the scripts with careful reference to the school's request and the mark scheme to decide whether the request should be upheld. If it is, the scripts are sent to a specially trained re-marker who re-marks the entire cohort. The re-marking may focus on certain questions or papers, in which case it is a partial re-mark, or it may be a full re-mark, in which case all components are re-marked. If the request is not upheld, the scripts are cleared and either sent for borderlining or sent straight back to the school.

The EMA's procedure for GRs in 2003 and 2004 only differed to take account of the separation and later re-combination of reading and writing papers. All other aspects of the review process from training to reporting remained the same. It is important to note again, however, that some schools may have approached GRs differently in 2004, because of the confusion caused by the difficulty in accessing results on the QCA database or the existence of incorrect marks on the database. Furthermore, borderlining had not taken place when schools received the scripts. In the past, borderlining took place before scripts were returned to schools, so GRs were requested on the basis that all pupils had been borderlined. In 2004, borderlining was re-scheduled to take place after the review process, which meant that pupils who would have had a level change due to borderlining would have been picked up in the review process instead.

Figures for the number of requests made and the number of reviews resulting in a change in level were available from 2003 which were used to compare GR requests in 2004.

Method

To estimate the number of requests resulting in a level change, the original marker's marks and the re-marker's marks had to be captured directly from a sample of scripts. The level change data from the Data Capture Agency (DCA) could not be used for this purpose because level changes attributable to incorrect data keying from an earlier stage of data capture could not be distinguished from those arising from GR re-marking. Opportunistic samples of scripts were keyed at the EMA, the sizes of which are given in Table 4. In this analysis, only changes to the reading level or writing level could be investigated, not final level changes. Whole pupil records could not be captured because the GRs available for keying were for either reading or writing. The mark data for the missing component on the QCA website could have been used to create whole pupil records if the missing component had been cleared or not requested for a GR. However, this was deemed risky because of the existence of incorrect marks on the database.

Table 4 Sample sizes for Group Review scripts for reading and writing in 2004

	Reading	Writing
Schools	35	28
Pupils	5,903	5,390

Results

The number of schools and pupils involved in GR requests made in 2003 and 2004 were very similar. The total number of individual schools which requested GRs in 2004 was 617, and in 2003, the figure was 592, as shown in Table 5. The number of individual pupils involved differed by less than 1,000. It should be noted that in 2004, the regime of component marking generated many more individual reviews because pupils for whom schools requested both reading and writing reviews (59,802 from 315 schools) had to be sent to two different reviewers, whereas in 2003, the same reviewer would have dealt with both components. It is also worth speculating whether the number of schools requesting a GR would have been less if borderlining had already taken place.

Table 5 The number of requests for Group Reviews made by schools and the number of pupils involved in 2003 and 2004

	2003	**2004***		
	Reading or writing or both	Reading or writing or both	Reading only	Writing only
Schools	592	617	438	494
Pupils	115,041	115,944	81,860	93,886

* Figures as at 1st October 2004.

The results of the initial review process in 2003 and 2004 differed: in 2003, the proportion of schools whose requests were not upheld was 33.9%; and in 2004, the proportion for reading was 37.0% and for writing it was 40.1%, giving an overall proportion of 38.6%. The number cleared in 2004 included some schools that were sent for borderlining. The decrease in the number of requests upheld suggests that schools sent in GR requests with weaker justifications than in previous years. Again, this may have been because they could not reconcile the results they saw on the database with those in the scripts.

The proportion of pupils whose reading or writing level changed, either up or down, as a result of a GR appeared to rise by 6.2 percentage points in 2004 compared with 2003, as shown in

Table 6. In 2003, of the 76,510 pupils involved in a GR, 21.7% of them saw a change in their level as a result of re-marking. In 2004, the level change rate in the combined sample of reading and writing re-marks was 27.9%. However, it is important to note that this proportion included level changes attributable to borderlining and is therefore an over-estimate of the actual number of level changes attributable to the review process.

Table 6 Proportion of level changes as a result of GR re-marking in 2003 and 2004

	2003		2004	
	Total number of re-marks	% of level changes	No. in sample	% of level changes
Reading or writing	76,510	21.7	11,293	27.9

Conclusion

The number of schools making a request for a GR remained very similar in both years, giving rise to the speculation that had borderlining taken place before the review process, the number of requests in 2004 would have been lower. The proportion of schools whose GR request was not upheld dropped in 2004, suggesting some schools mis-targeted their requests and some used the GR request as a means of borderlining. The proportion of pupils whose GR re-marks resulted in a level change rose in 2004, but because of the presence of pupils who would have had a level change through borderlining, it is difficult to attribute this increase solely to marking errors. No records of the proportion of level changes as a result of borderlining were ever kept by markers so it is difficult to assess the impact the un-borderlined pupils would have had, but it is highly likely some impact would be seen in the level changes in GRs. This suggests that the 6.2 percentage points increase in level changes observed in the sample was not of great significance.

5. REFERENCES

Schermbrucker, J. (2004) *External marking of key stages 2 and 3 national curriculum tests and year 7 progress tests in 2004: Analysis of 2003 Reviews Statistics – Key Stage 3 English.* Internal report of External Marking Agency, AQA.

Lucy Royal-Dawson
AQA Senior Research Officer
11th October 2004

APPENDIX

STANDARDISATION SAMPLE
READING

In 2003, markers were required to send their team leader four reading scripts and six Shakespeare reading scripts for standardisation purposes. In 2004, they had to send ten of each.

Table 1 Standardisation sample: reading component

a) Rate of agreement in marks as a percentage of the total number of pupils

		2003			2004		
	Max mark	No. of pupils	0 marks different	1 mark different	No. of pupils	0 marks different	1 mark different
Reading	32	212	12.7	33.4	490	16.1	51.3
Shakespeare reading	18	318	17.0	50.6	489[2]	24.1	57.7

b) Mean pupil AMD

	2003			2004		
	No. of pupils	Mean AMD	SD	No. of pupils	Mean AMD	SD
Reading	212	2.69	2.03	490	1.79	1.46
Shakespeare reading	318	1.80	1.42	489	1.66	1.57

WRITING

In 2003, markers were required to send five writing scripts and six Shakespeare writing scripts for standardisation purposes. In 2004, they had to send ten of each.

Table 2 Standardisation sample: writing component

a) Rate of agreement in marks as a percentage of the total number of pupils

			2003			2004		
	Strands	Max mark	No. of pupils	0 marks different	1 mark different	No. of pupils	0 marks different	1 mark different
Writing	SSP	8	285	24.2	70.5	380	37.6	83.7
	TSO	8	285	17.5	70.8	380	37.1	78.4
	C&E	14	285	19.3	53.4	380	26.3	63.2
Shakespeare writing	SSPTO	6	341	36.1	85.1	380	45.8	86.6
	C&E	10	341	28.2	64.0	380	30.8	67.9
	Spelling	4	341	49.9	94.8	380	50.5	93.4

b) Mean pupil AMD

	2003			2004		
	No. of pupils	Mean	SD	No. of pupils	Mean	SD
Writing	285	4.06	2.56	380	3.17	2.49
Shakespeare writing	341	2.66	1.82	380	2.37	1.67

[2] The number of pupils in the tables in the Appendix is not always the number of pupils in the sample multiplied by the number of markers because some marks were illegible or missing on the forms.

FIRST SAMPLE

READING

The reading first sample in both 2003 and 2004 consisted of ten scripts of both components.

Table 3 First sample: reading component

a) Rate of agreement in marks as a percentage of the total number of pupils

		2003			2004		
	Max mark	No. of pupils	0 marks different	1 mark different	No. of pupils	0 marks different	1 mark different
Reading	32	530	33.8	68.4	484	30.0	68.7
Shakespeare reading	18	530	59.1	84.9	477	55.8	78.9

b) Mean pupil AMD

	2003			2004		
	No. of pupils	Mean	SD	No. of pupils	Mean	SD
Reading	530	1.27	1.39	484	1.21	1.15
Shakespeare reading	530	0.61	0.92	477	0.78	1.10

WRITING

The writing first sample in both 2003 and 2004 consisted of ten scripts of both components.

Table 4 First sample: writing component

a) Rate of agreement in marks as a percentage of the total number of pupils

			2003			2004		
	Strands	Max mark	No. of pupils	0 marks different	1 mark different	No. of pupils	0 marks different	1 mark different
Writing	SSP	8	560	67.5	94.3	380	68.7	95.5
	TSO	8	560	71.4	95.5	380	64.7	95.5
	C&E	14	560	63.2	90.9	380	55.0	87.9
Shakespeare	SSPTO	6	560	70.7	98.2	380	71.1	94.8
Writing	C&E	10	559	67.1	93.1	380	57.4	87.7
	Spelling	4	559	74.2	99.6	380	77.4	97.7

b) Mean pupil AMD

	2003			2004		
	No. of pupils	Mean	SD	No. of pupils	Mean	SD
Writing	560	1.20	1.33	380	1.37	1.43
Shakespeare writing	559	0.98	1.08	380	1.17	1.28