Qualifications and Curriculum Authority

Research into marking quality

Studies to inform future work on national curriculum assessments

March 2009

QCA/09/**4042**

Contents

Introduction

This paper has been produced to inform the regulator of the work undertaken by the Qualifications and Curriculum Authority (QCA) and the now disbanded National Assessment Agency (NAA), in relation to the quality of marking.

Purposes of this paper

The purposes of this paper are:

- to outline the research activities undertaken as part of the programme of work to improve the quality of marking in national curriculum tests
- to identify the impact of research findings on the national curriculum tests programme by setting out the actions taken in response to these findings
- to identify the issues with marking quality and explain the further work that QCA will be undertaking.

Background

Improving the quality of marking formed part of NAA's remit from its launch in April 2004 and it continues to be central to the work of QCA. Since 2004, a number of research projects focused on improving the quality of marking have been undertaken by NAA itself and through commissioned external agencies. This research has spanned a range of tests and qualifications – from national curriculum tests at key stages 2 and 3, to General Certificate of Secondary Education (GCSE) and Advanced level (A level) qualifications. However, the basis of the research is to inform future work on national curriculum tests.

The research has always had a strong focus on operational issues, and this report sets out the actions taken as a result of the findings of each research study. It also outlines areas that require further investigation by QCA.

Outline of the issues

To underpin its programme of work to address the quality of marking, NAA commissioned the Assessment and Qualifications Alliance (AQA) to review the literature on marking reliability (Meadows & Billington, 2005). This review had four main purposes:

- conceptual clarification and mapping of existing literature on marking reliability
- identification of sources of bias in marking

- identification of the effect of different types of assessment on marking reliability
- identification and evaluation of remedial measures to detect/correct unreliable marking.

These are discussed in detail below.

Conceptual clarification and mapping of literature on marking reliability

Meadows & Billington (2005) clarified the concept of marking reliability by identifying the range of definitions and forms of marking reliability, and outlining the relationship between reliability and validity. They then identified and distinguished between different measures of reliability: test-retest; split-half; internal consistency; and alternate-form. They also identified the different types of inter-marker reliability – consensus, consistency and measurement estimates.

This review also mapped national and international literature on marking reliability, providing NAA with a clear overview of work that had already been undertaken in this field and identifying areas requiring further investigation. This review set out, at a high level the major sources of bias in marking identified in the literature.

Impact

Meadows & Billington's (2005) identification of the different types of inter-marker reliability – consensus, consistency and measurement estimates – will help QCA establish of the most appropriate measure of marking reliability. For national curriculum tests, in line with general qualifications, the correct mark was defined as the mark which the most senior marker would have given a piece of work. In the light of thinking about new and flatter structures for marking teams in testing programmes such as single level tests – more appropriate for an increasingly professionalised marking community – the impact of using alternative definitions will be explored through modelling. The findings will inform QCA's thinking on the most appropriate definition of the correct mark in new marking contexts in which traditional marking structures may not be appropriate, such as on-demand and computer-based assessments.

Sources of bias in marking

The main high-level sources of bias in marking which Meadows & Billington (2005) reviewed were:

- **Contrast effects**: this is when the marks awarded to a script are influenced by the standard of preceding scripts.

Impact
For operational purposes – as marking is carried out on paper – it has not been possible to investigate this potential source of bias in the context of national curriculum tests. The introduction of on-screen marking, in any testing programme, could allow the splitting of scripts for a school between different markers and so enable the impact of reducing any possible contrast effects to be evaluated.

- **The text (handwriting) of the script**: the review of the literature concluded that, where scripts were marked by experienced examiners using clear mark schemes, the possibility of bias arising from the text of the script was reduced (Baird, 1998).

Impact
Given that Baird (1998) suggested that there was little research evidence to suppose that there would be significant bias arising from the text of the script**,** in the context of national curriculum testing, no work in this area is currently planned.

- **The candidate**: literature on the potential for gender/ethnicity bias in marking was reviewed.

Impact
Given that Meadows & Billington (2005) concluded that, where there were clear mark schemes and thorough marker training/monitoring, it was unlikely that there would be bias arising from the marking, no work in this area is currently planned.

- **The examiner:** studies which had investigated the impact of examiner background and examiner traits on marking reliability were reviewed.

Impact
The NAA commissioned a further study (Meadows & Billington, 2007), investigating the impact of examiner experience and character traits empirically. This study found that there were no justifiable reasons for taking examiner background into account when making judgements about recruitment. Following the announcement that national curriculum testing will not longer be statutory at key stage 3, QCA is permitting experienced markers for key stage 3 to be recruited for key stage 2 marking in 2009.

However, priority is given to those with qualified teacher status with experience of key stage 2 teaching. This practice will be reviewed at the end of the cycle.

Effect of different types of assessment on marking reliability

Meadows & Billington (2005) outlined the evidence on the impact of subject on marking reliability, finding that much of the impact arose from the particular question formats used for particular subjects. This issue is considered further in section 5, where marking reliability for national curriculum tests in English, mathematics and science is discussed in more detail.

In general terms, Meadows & Billington (2005) noted that highest marking reliability is achieved with multiple-choice items. However, this needs to be balanced against validity considerations.

Impact

At present, there are no plans to make changes to the format and structure of national curriculum tests at key stage 2. However, the implications of this are being carefully considered in the development of single level tests, particularly with respect to English reading and writing, to ensure that items are optimally valid and the marking of them reliable.

Meadows & Billington (2005) set out the evidence on the impact of mark schemes on marking reliability, suggesting that marker input into the development of the mark scheme could improve reliability.

They set out the evidence on the relative merits of holistic and analytic scoring, arguing that there is a place for both, but that analytic scoring may be thought to be more useful for accountability purposes.

Impact

Markers are involved from an early stage in the development of national curriculum tests and this will continue. Although there are no plans to change the format of mark schemes in national curriculum testing at present, QCA is undertaking research into mark schemes and mark scheme development in the context of single level tests.

Remedial measures to detect/correct unreliable marking

Meadows & Billington (2005) outlined the methods currently in use to detect unreliable marking, and then identified the ways in which such marking might be detected more reliably once scripts were marked on-screen. Meadows & Billington (2005) identified a number of advantages in relation to on-screen marking in terms of detecting unreliable marking/markers. These include the possibility of seeding items, blind double marking and real-time monitoring of markers.

Impact
QCA is committed to introducing on-screen marking within its testing programmes because of the benefits for quality of marking. This is already taking place in the SLT pilot.

Meadows & Billington (2005) pointed out that there is relatively little research on potential remedial measures to address unreliable marking. The most common method – mark adjustment, where scores for a marker are adjusted up if they are found to be marking harshly or down if they are found to be lenient – is not feasible for current national curriculum tests, because scripts are returned to schools with the marker's original mark on them. As more marking is carried out on-screen, this option could be examined further. Were marking adjustments to be considered, however, a programme of research would need to be undertaken to address issues surrounding this method – such as the impact of applying uniform adjustments to a single marker.

Elements relating to marking reliability

In this section, we consider in more detail the elements relating to marking reliability, as identified by Meadows & Billington (2005).

Sampling

In both national curriculum tests prior to 2008 and general qualifications, the marking of markers/examiners was monitored by a senior marker/examiner. This was carried out by sampling, which involves the senior marker/examiner reviewing samples of live marking by a marker/examiner. The measure of the difference between the marks of the senior marker/examiner and those of the marker/examiner improves with larger samples, but there was no clear evidence about the optimum number of scripts which should be reviewed in order to most effectively and efficiently evaluate the marking of a marker/examiner. The NAA commissioned a research project (Al Bayatti & Jones, 2005), to establish a way of determining

the minimum sample sizes required to detect specific differences (in terms of marks) between the marks of the senior marker and the marker. The study also showed that the number of scripts required to identify markers with errant marking varied according to teaching/marking experience; the more experienced the marker, the smaller the sample size required.

Meadows & Billington (2005) provided evidence that suggested that the kind of remarking that took place when scripts were sampled led to artificially high estimates of marking reliability. This was because in instances where the original marker's mark was on a script, the senior marker tended to slightly modify the marks of the marker, rather than remark the work (McVey, 1975).The research also showed that, when marked scripts were given to senior markers to remark, but the marks were hidden, there was evidence of greater variation between senior markers' and markers' marks (WJEC, 2004).

In national curriculum testing, an additional issue has been that there is traditionally an element of markers self-selecting the scripts that make up their samples and the first sample taken during live marking was linked to standardisation, reducing the number of occasions on which markers were monitored.

Impact

In 2008 quality assurance throughout the marking window was separated from standardisation at the start of the marking window and NAA introduced a new system of benchmarking to replace the sampling process. In benchmarking, markers are presented with a number of scripts for which a correct mark has been agreed in advance. Markers were then expected to mark these scripts, and their marks had to be within an agreed tolerance if they were to be permitted to continue marking. The correct mark is not visible to the marker. Such benchmarking removed the sample self-selection and variability between senior markers that posed a risk to reliability. Furthermore, because it took place online, such benchmarking could be carried out more frequently than previous sampling activities.

In 2009 QCA intends to maintain the concept of benchmarking, although due to the lateness of the contract award, there is no time to develop an online system for delivery. As a result, quality assurance against the correct marks will be conducted using a paper-based system, and will therefore only occur twice during the marking period (this is in addition to standardisation, which will take place at the start of the process). The intention is to increase the frequency of quality assurance checks in future, particularly with the introduction of on-screen marking.

Double marking

A study was commissioned from AQA (Fearnley, 2005) on double marking. The approach of the study was pragmatic, recognising that, while double marking might be desirable in principle, within the current system in which there can be difficulties recruiting markers and time-pressure is already felt by senior markers, it might be difficult in practical terms. There are also logistical implications for moving scripts between several markers. Whether or not double marking should be introduced is, therefore, a decision that needs to be made through careful consideration of not only the benefits, but also of the risks and costs. For this reason, Fearnley (2005) investigated whether a system of double marking could be introduced which would not put intolerable strains on senior markers.

There are also costs involved in the process of double marking. Fearnley (2005) investigated whether double marking when a second marker had access to the first marker's marking was preferable to double marking using cleaned scripts. In a paper-based system, producing clean scripts for double marking could be costly and time-consuming.

Finally, Fearnley (2005) investigated whether double marking could be targeted, to ensure scripts that were at the highest risk of being marked unreliably could be double marked.

The study found that double marking was more effective on clean scripts, but the major finding was that improvements in marking reliability arising from double marking were very small and it was therefore judged not to be cost effective.

Impact

The QCA will only consider double marking following the introduction of on-screen marking, when blind double marking will be possible and may be cost and time effective.

Question papers and mark schemes

Meadows & Billington (2005) set out the evidence in relation to the impact of subject and item format on marking reliability. They showed that, as expected, closely identified questions demanding definite answers were associated with high levels of marking reliability. Evidence for this in the context of national curriculum tests, had been found by Baker et al (2006) in an investigation into the reliability of marking for key stage 3 mathematics tests, and it would be plausible to expect lower levels of marking reliability for a subject like English. Newton (2006), discussing his finding that GCSE English had lower marking reliability than GCSE mathematics, had argued that the difference in marking reliability between question types was inevitable as long as particular assessment formats were valued.

For national curriculum tests in English, for validity reasons, a range of response formats (short response, constrained response, extended response) were valued and Baker et al (2008) showed that these formats gave rise to differing levels of marking reliability (defined as agreement with the correct mark). It should be borne in mind, when interpreting the findings of this study, however, that it was carried out as part of an international study involving the marking of national curriculum tests in Australia, using Australian teachers of English. Therefore the findings might not be generalised without difficulties to marking carried out by English teachers of English.

Baker et al (2008) found that, for national curriculum tests of English Reading, items that were highly constrained – that is, low tariff questions, usually of one-mark value, in which the acceptable answers were prescribed clearly by the mark scheme – had marking reliabilities for individual items ranging between 87.13% and 98.98%. For such items, marker agreement should be very high indeed, approaching 100%. For mid-constrained items – that is, medium tariff questions – the marking reliability ranged between 75.28% and 91.30%, and for open-ended response items the marking reliability ranged between 48.83% and 62.32%.

Even if these findings were partially due to the differences in educational/cultural background of the Australian markers, there is a basic issue to be investigated. While recognising the importance of a variety of response formats for validity reasons, NAA was concerned about these potentially variable and sometimes low levels of marking reliability.

Impact

There are no plans to change the format or structure of the national curriculum tests. However, within the context of single level tests, QCA is investigating the possibility, for English tests, of using item types that can be more reliably marked without a loss of validity.

Turning from the question papers to the mark schemes, Meadows & Billington (2005) investigated the impact of different mark schemes, setting out the evidence for the reliability of holistic and analytic marking. In the context of national curriculum tests, they cited evidence that analytic marking of key stage 3 English scripts could, in some circumstances, lead to depressed marks and more erratic marking (UCLES, 2000). However, they also pointed out that, where holistic scoring is used, it can be difficult to be certain what criteria markers are using when allocating their marks, and it may be that markers are using criteria that are not construct-relevant – that is not relevant to the subject being tested.

Impact
As indicated above, given that there are no plans to change the model of assessment in national curriculum tests, QCA is undertaking research into mark schemes and mark scheme development in the context of single level tests.

Marker selection

The selection of markers for UK national examination and testing systems is largely a matter of custom and practice. Examiners/markers are generally required to have suitable academic qualifications and recent, relevant teaching experience. In order to respond to demands for a greater number of examiners/markers, while maintaining high standards of marking quality, NAA identified the need to establish empirically supported examiner/marker recruitment and selection practices. A study carried out in 2004 suggested that there might be no significant difference, in terms of marking reliability, between experienced key stage 3 English markers, experienced English teachers with no external marking experience, recent PGCE graduates in English with teaching experience gained on teaching practice, and recent BA graduates in English with no teaching experience (Royal-Dawson, 2004). The study recommended that the criterion of teaching experience could be relaxed to allow markers with graduate-level subject knowledge to mark key stage 3 English tests.

Meadows & Billington (2005) had found that the research evidence on the impact of marker background and marker traits on marking reliability was far from conclusive. Given this uncertainty and the relatively small scale of the 2004 study, a further study was commissioned (Meadows & Billington, 2007). This study focused on marking in GCSE English and involved a higher number of participants (359 in total). It explored differences, in terms of marking reliability, between participants from different educational, teaching and examining backgrounds: experienced examiners (with high subject knowledge and teaching experience), PGCE English undergraduates (with high subject knowledge and some teaching experience), English undergraduates (with high subject knowledge and no teaching experience), and non-English undergraduates (with low subject knowledge and no teaching experience). It also explored the value of using measures of personality and attitude as predictors of marking reliability for these participants.

The findings were that examiners marked more reliably than both groups of undergraduates, indicating that subject knowledge and some experience of teaching is important for marking reliability. PGCE students marked less reliably than experienced examiners, but better than both groups of undergraduates. The study also found that examiner/marker training had the effect of compressing the distribution of marks awarded by participants, despite its function

being to stretch that range to avoid compression of the final mark distribution and hence of the grade boundaries. Further investigation into the reasons for this was recommended.

Impact
The recruitment criteria for national curriculum tests prioritise experienced markers over all other groups, with experienced teachers who have no previous experience of external marking as second preference before PGCE students. Given the findings of this study, QCA has no plans to amend these criteria.

Marker training

A preliminary study of the training models used for key stage 2 and key stage 3 English was also carried out (Muallem, 2006). This study identified time pressures during the marker training day as a major issue for both key stage 2 and key stage 3 and argued that the proportion of time spent on administrative training was perceived to have a negative impact on the time available for subject-specific training. Further, the quality of administrative training varied significantly between teams. Muallem (2006) also argued that more emphasis should be placed on the interpersonal skills of senior markers, rather than simply on marking skills, as their ability to train and support markers was key.

The findings of Muallem (2006) finding were validated by FreshMinds (2007), whose study into alternative models of marker training recommended a blended approach to training, with elements such as administrative training delivered online and additional, softer skills-based training for senior markers and team leaders.

Impact
In the context of national curriculum tests, the issues identified by Muallem (2006) and FreshMinds (2007) were addressed in 2008 and all administrative training was delivered online, freeing up more time for subject-specific training and also allowing markers to work through the administrative procedures at their own pace, revisiting particular issues if they were uncertain or unclear. In 2009, administrative training will continue to be undertaken outside of the marker training meetings.

With the appointment of the new marking contractor for 2008, NAA judged that it was important to gain an understanding of the perspectives of the marking community. A study was commissioned (Fitzgerald, Hughes & Goodwin, 2007) on marker attitudes. The aim of this study was to elicit from the marking community their perceptions of the marking process. It

was hoped that this would give NAA information about the kinds of concerns the marking community would have at a time of rapid change, enabling them to put in place systems to support and reassure markers.

Impact
The problems in 2008 mean that greater attention needs to be paid to the marker experience, to ensure they are able to focus on marking reliability rather than on administrative issues.

Having outlined and considered the research work which NAA undertook with respect to the quality of marking, we now turn to the work which NAA undertook to attempt to baseline the quality of marking in the context of national curriculum tests. We also outline QCA's proposals for the future.

Baselining the quality of marking, and proposals for the future

National curriculum test marking is carried out by an external contractor: the marking agency. It is the responsibility of the marking agency to ensure that marking is reliable and to put in place quality assurance systems which enable errant marking to be detected and appropriate interventions to be made. It is the responsibility of QCA to ensure that this quality assurance takes place and to evaluate its effectiveness. All QCA work on national curriculum tests is monitored by Ofqual

The NAA began its programme of work on the quality of marking with the recognition that there was no clear baseline for marking reliability in relation to national curriculum tests. There were also no established benchmarks against which any interventions might be evaluated. For these reasons, NAA began by attempting to establish a measure of the current levels of marking reliability for national curriculum tests. To do this, NAA undertook a small-scale exercise, the purposes of which were to identify, as far as possible, a benchmark estimate of marking reliability for national curriculum tests and to generate hypotheses about the potential sources of any unreliability. A suite of small-scale studies was carried out in 2006 and 2007, in which marking for key stage 3 science, English and mathematics was examined. Key stage 3 was chosen to start with because the nature of the tasks, with more open ended responses, was likely to lead to greater inconsistency.

There were issues to do with the design and implementation of the science and English studies that meant that the findings might not be immediately generalisable. First, the sample

of participants was not strictly random. Markers were approached randomly, but the decision about whether or not to participate was made by each individual. Second, participants were aware that this was a study and not operational marking, so could have no impact on the pupils whose work was marked, nor on their own records as markers (marking was undertaken anonymously). This might have predisposed markers to take rather less care with the marking than they normally would. Third, the marking took place at the end of the marking period, so there might well have been a fatigue effect.

In spite of these caveats, the fact that the studies indicated that all the markers were marking at the low end of acceptable marking reliability suggested that the quality of marking was an issue. For these reasons, the findings from these studies need to be treated as indicating that there is an issue to investigate further and that there appears to be a more significant issue in English.

Mathematics

The findings for key stage 3 mathematics (Baker et al, 2006) showed very high levels of inter-marker agreement (Cronbach's Alpha, 0.95) and almost perfect intra-marker reliability (Kappa Coefficient, 0.99). There was no evidence that item type or response format had any impact on marking reliability. However, the number of markers involved was small.

Science

For key stage 3 science in 2006 the analysis indicated that 86.6% of pupils whose work was marked by a marker would have been awarded the same level had their work been marked by a senior marker.

For key stage 3 science in 2007 the analysis indicated that 87.4% of pupils whose work was marked by a marker would have been awarded the same level had their work been marked by a senior marker. The difference between the findings from 2006 and 2007 are not statistically significant ($p \leq 0.05$).

English

For key stage 3 English reading in 2006 the analysis indicated that 66.2% of pupils whose work was marked by a marker would have been awarded the same level had their work been marked by a senior marker. For key stage 3 English reading in 2007 the analysis indicated that 70.7% of pupils whose work was marked by a marker would have been awarded the same level had their work been marked by a senior marker.

For key stage 3 English writing in 2006 the figure was 56.4% and in 2007 it was 55.4%. Again, differences are not statistically significant ($p \leq 0.05$).

Analysis of the benchmarking data from 2008

The process changes in 2008 made a large amount of quality assurance data available to the NAA for the first time. The NAA commissioned the University of Bristol (Royal-Dawson, Leckie and Baird, 2009) to produce misclassification statistics from the standardisation and benchmarking data in 2008.

The authors were asked to analyse the data in terms of classification consistency, as this would be a concern for key stakeholders. It was recognised, by both the authors and the NAA, that this would pose challenges and Royal-Dawson, Leckie & Baird (2009) made clear throughout their report the caveats that needed to be borne in mind.

One of the major issues was that the data analysed had been generated from an ongoing quality assurance process, rather than within the context of a designed study, so inferences had to be drawn with caution. A study designed for the purposes of establishing classification consistency would require a sample of scripts on mark points mirroring the national distribution, but such scripts were not selected for standardisation/benchmarking purposes (because this is not a requirement for the quality assurance process) and the Bristol study noted that the potential impact of scripts on mark points near a threshold on its findings, with small variations in marking potentially magnifying apparent misclassification. Conversely, if scripts under-represent proximity to the cut-scores, then the apparent misclassification rates would be an under-estimate. For that reason, inferences about the proportion of pupils who were correctly classified need to be interpreted not as claims about classification consistency for pupils nationally in 2008, but as indicative of the quality of marking within the quality assurance process.

Table 1 below sets out the overall rate of agreement between a marker's level and the true level for every candidate in the standardisation and benchmarking sets for each test and for the two components of English at key stage 3.

Table 1: Classification rates and one level difference for all tests at standardisation and benchmarking

	Standardisation			Benchmarking		
	Exact agreement	Markers +1 level	Markers -1 level	Exact agreement	Markers +1 level	Markers -1 level
Key stage 2						
English reading	74.1%	24.8%	1.1%	87.4%	10.4%	2.1%
English writing	90.4%	3.5%	6.1%	77.9%	10.3%	11.8%
Mathematics	99.9%	0.1%	0%	99.5%	0.5%	0.1%
Science	99.0%	0.1%	0.9%	96.6%	2.5%	0.9%
Key stage 3						
English reading	78.4%	12.6%	8.7%	66.6%	22.0%	10.8%
English writing	66.5%	17.2%	14.5%	63.7%	16.1%	17.6%
Mathematics	88.5%	3.2%	8.3%	98.0%	0.3%	1.7%
Science	97.9%	1.1%	1.0%	94.5%	3.2%	2.2%

Classification rates for the English components were lower than those for both Maths and Science at both Key Stages, as would be expected given that they consist of items which tend to yield higher mark differences, and for which higher levels of discrepancy were tolerable. Differences in levels tended to be on the generous side, with markers' levels being one level higher than the true mark, particularly so for English Reading. Given that Key Stage 3 Maths consisted mainly of one or two mark items, the classification rate in the standardisation sets of 88.5% may be considered surprising, but further analysis indicated that the disagreements were almost entirely attributed to two candidates whose true marks was on or very close to a level threshold.

The absence of studies based on operational marking from previous years meant that comparative judgements could not be made in this context, but the authors of the study were able to make judgements based on the levels of marking quality which they had noted in other examination contexts, again emphasising the indicative nature of the conclusions. The authors concluded that, on the basis of the analyses which they had carried out and their knowledge of the quality of marking in other contexts," Although, given the foregoing caveats, these findings were not a robust measure of the 2008 marking quality of national curriculum

tests, neither were the figures a cause for particular concern" (Royal-Dawson, Leckie & Baird, 2009 p1).

Impact

The solution proposed and implemented by the new marking agency in 2008 was:

- ensuring more consistency in training by reducing the number of presenters
- changing the standardisation process, such that markers were checked centrally, against an agreed national standard, rather than checked by their team leader
- checking marking accuracy more frequently, at four rather than two intervals during marking. These checks are also carried out against the national standard.

To further increase the reliability of the reporting of outcomes:

- markers recorded marks electronically, reducing the likelihood of clerical errors in the transcription and totalling of marks
- the process of assigning levels was automated, reducing the possibility of clerical errors.

Although the problems in 2008 are well documented, it would appear that the processes used to ensure marking quality were robust. However, following the termination of the contract and the award of the contract to a new supplier at such a late stage in the year, it will not be possible to fully replicate these processes in 2009. Nonetheless, QCA is trying to preserve the main principles where possible. These include:

- For English, training will be delivered in the same way as in 2008. Maths and Science will revert to a small group training model.
- Standardisation will be undertaken against a national standard although it will not be carried out online.
- The benchmarking process will be carried out on paper. As a result, marking will be checked on fewer occasions, but with the same number of scripts.
- Markers will again record marks on paper marksheets, but the assigning of levels will be done automatically, as well as by markers, to act as a check.

Conclusion

The NAA has undertaken a programme of work, outlined above, to conceptualise marking reliability, develop and trial a range of approaches to measuring it, and begin to work with partner agencies and other stakeholders on making carefully targeted improvements. The QCA is now well placed to continue this work in partnership with the new marking agency. Issues relating to marking reliability are now embedded in the evaluation processes for national curriculum tests and single level tests and, over time, these processes will be reviewed and refined.

References

Al Bayatti & Jones (2005). *The effect of sample size on increased precision in detecting errant marking*. AQA: Manchester.

Baird J (1998) *What's in a name? Experiments with blind marking in A-level examinations.* Educational Research, v40, n2, p191–202.

Baker EL, Ayres P, O'Neil H, Choi K, Tettey M and Sylvester R (2006) *KS3 Mathematics Marker Study in Australia: Report to the National Assessment Agency of England.* University of Southern California: Sherman Oaks, CA.

Baker EL, Ayres P, O'Neil H, Choi K, Sawyer W, Sylvester RW & Carroll B (2008). *KS3 English Marker Study in Australia: Final Report to the National Assessment Agency of England.* University of Southern California: Sherman Oaks, CA.

Fearnley A (2005) An Investigation of Targeted Double Marking for GCSE and GCE. AQA: Manchester.

Fitzgerald, Hughes & Goodwin (2007). *National Curriculum English Test Markers' Attitudes to Training*. Edexcel: London.

FreshMinds (2007) *Marker Training Models: exploring alternative delivery structures*. FreshMinds: London.

McVey P J (1975) *The errors in marking examination scripts in electronic engineering.* International Journal of Electronic Engineering Education, v12, p203–216.

Meadows M & Billington L (2005). A Review of the Literature on Marking Reliability. AQA: Manchester.

Meadows M & Billington L (2007). NAA Enhancing the Quality of Marking Project: Final Report for Research on Marker Selection. AQA: Manchester.

Muallem D (2006) *Evaluation of KS2 and KS3 English Marker Training Models.* NAA.

NAA (2005). *Marking Quality Index: English and Science*. Internal paper.

Newton P (1996) *The reliability of marking General Certificate of Secondary Education Scripts: Mathematics and English.* British Journal of Educational Research, v22, n4, p405–420.

Royal-Dawson L (2004) *Is teaching experience a necessary condition for markers of Key stage 3 English?* AQA: Guildford.

Royal-Dawson L, Leckie G & Baird J (2009). *Marking reliability of the 2008 national curriculum tests*. University of Bristol

University of Cambridge Local Examinations Syndicate (UCLES) (2000). *Key stage 3 English – A study of marking reliability which investigates three different methods of maintaining consistency between markers.* A report produced for the Qualifications and Curriculum Authority.

Welsh Joint Education Committee (WJEC) (2004) *CMI/CMS Pilot Evaluation.* Welsh Joint Education Committee.