Centre for Education Research and Policy

Identifying Errant Markers: Quality Assurance Systems in an E-Marking Environment

Anne Pinot de Moira

The exploitation of innovative technology in the administration of high stakes national qualifications, has provided the opportunity to develop robust quality assurance systems to monitor and improve marking reliability. This paper uses Monte Carlo simulation to explore two simple models of quality assurance: a hierarchical seeded system and a peer-pair double marked system. The models suggest that item marking minimises the effect of systematic differences between markers but the extent to which this improves reliability at a paper level is dependent upon the number of items on a paper. Beyond the positive effect of dividing a single paper among different markers, it is concluded that a quality assurance system which includes any element of sampling has little potential to influence marking reliability *directly*. It becomes a system for identifying errant markers for retraining. It is shown that, for a given set of quality assurance parameters, the peer-pair double marking system is more likely to identify an errant marker than the hierarchical seeded system.

Keywords: hierarchical seeded quality assurance system; peer-pair double marked quality assurance system; marking tolerance; e-marking.

The Practicalities of E-Marking

The potential to use innovative technology in education and assessment has been recognised for many years. Initially the focus was on teaching and learning in the classroom (see, for example, Sheingold, Kane, Endreweit, & Billings, 1981; Sheingold & Tucker, 1990) but the value for assessment provision and administration was quickly embraced (Means *et al.*, 1993). Since its infancy at the turn of the century, the use of technology in the provision of high stakes national examinations in the United Kingdom has proliferated. This has been particularly apparent in the area of on-screen marking, or e-marking, where candidate responses are presented to a marker electronically and that marker never has sight of the original paper version.

With the rapid advances has come a need to ensure the maintenance of assessment reliability and validity. Studies have seemed to confirm that e-marking provides a satisfactory substitute for paper-based marking, although findings have been limited by the experimental design. The validation of e-marking systems has been necessarily coincident with the development of new technology. Therefore the extent to which these studies have successfully replicated a live marking environment is restricted and, moreover, the published evaluation of operational quality assurance systems is largely absent. Nevertheless, Fowles (2002) showed that there was a close relationship between marks awarded by chemistry markers using conventional paper-based marking and those awarded using e-marking. Further evidence was presented in Powers, Farnum, Grant, & Kubota (1997), Whetton & Newton (2002) and Sturman & Kispal (2003), all of whom reported a high correlation between the outcomes of conventional and on-screen marking.

Once assured that marking is unimpaired by the electronic environment, the advantages of e-marking are understood to be numerous. Whetton & Newton (2002), Hudson (2009) and Williams & van Lent (2002) all highlighted potential benefits. They can be paraphrased thus:

1) marking would be speeded up, as papers would spend less time in transit;
2) data collection would become entirely automated, eliminating the inevitable errors that creep into manual processes;
3) marking might become more accurate, as markers could focus upon specific questions, termed items, rather than having to mark all items from an examination;
4) items that do not require subject matter expertise to mark could be marked by non-experts, thereby using human resources far more effectively;
5) marking might become less biased, as the introduction of randomly presented item marking would remove halo effects;
6) individual marker idiosyncrasies might be offset because multiple markers would contribute to the overall score for a paper;
7) the complete anonymisation of work might make marking less biased; and
8) marking might become more reliable and valid as a result of the introduction of real-time quality assurance systems.

In terms of marking reliability, the cumulative effect of these benefits amounts to the reduction in systematic errors. Time constraints are alleviated, clerical errors are eliminated and expert markers' focus is directed to the items where their knowledge is best used. Furthermore, by dividing items among markers so whole-paper marking[1] is a thing of the past, individual marker idiosyncrasies are said to be offset. However, the extent to which any of these benefits manifests themselves has been little explored.

The advantages to be gained by removing the more clerically arcane parts of paper-based marking systems are dependent upon the efficiency of the new systems. Such new systems are commercially sensitive and it is, therefore, unsurprising that the benefits have not been widely reported. Nevertheless in 2007, Taylor showed that, following the transfer of examinations to e-marking, there was a small drop in the number of post-result enquiries which culminated in a mark change. Sadly, this pattern was not replicated in a similar study the following year (Taylor, 2008).

Evidence-based research into the reliability and validity of e-marking has been largely restricted to the aforementioned comparisons with a paper-based approach. Some researchers have considered the benefit gained from the use of non-expert markers; a facility made easier when item marking with an e-marking system (see Meadows & Billington, 2005 pp 30-35 for a full discussion). However, even though simple statistical theory should tell us that item marking is preferable to whole-paper marking in terms of arriving at some *true mark*, no empirical evidence has yet been presented. If it is known that the error associated with scoring each item on a paper is centred around zero and the items can be considered independent with identical distributions of error, then according to the central limit theorem, the error associated with the scoring of the paper (the sum of the items) should be approximately normally distributed with a mean of zero. The assumption of independence is clearly not met when one marker scores every item on a paper but, when item marking is introduced, so too is a level of independence. The error distributions for each item on a paper are

[1] Whole-paper marking is defined as all items on one paper being marked by a single marker in the order in which they were presented within the paper.

unlikely to be identical but the extent to which this affects the impact of using multiple markers is little explored. Furthermore, there is no evidence to support the advantages of real-time quality assurance systems for the standardisation and monitoring of markers. Indeed, whether in the context of e-marking or conventional marking, there exists little reported research into the mechanics of implementing quality assurance systems during the process of marking.

Costs and time constraints dictate that extensive multiple marking of responses is not viable in high stakes UK national examinations. Nevertheless multiple marking is still widely regarded within some sectors of the education community as the most effective method of ensuring the final mark awarded is consistent with the mark scheme (see for example Lucas, 1971; Meadows & Billington, 2005; Pilliner, 1968; Qualifications and Curriculum Authority (QCA), 2002; Wood & Quinn, 1976). This view of quality assurance is not one that is shared within the world of manufacturing. Indeed when Deming (2000) laid out his fourteen key principles for quality control, his third principle suggested that companies should "cease dependence on mass inspection". He continued by saying that "inspection does not improve quality, nor guarantee quality" (pp.28-29); an idea that will be returned to later within this paper. However, he was clear to point out that there is merit in "the inspection of small samples of product to achieve or maintain statistical control" (p.29).

While the marking of national assessments differs in many ways from manufacturing, not least because of a human subjective element in process, e-marking systems lend themselves to a more mechanistic style of maintaining control of marking quality. For many years sample double marking has been used within the UK examination boards in order to assess the consistency and accuracy of individual markers. Decisions on sampling rates and acceptance criteria have, however, been largely pragmatic and certainly rarely the subject of peer review journal articles. Indeed, in his international review of marking quality assurance procedures, Lamprianou (2004) demonstrated that many examination systems in Europe, Australasia and within the United States of America, favoured double marking. Notable exceptions, were the West Virginia Educational Standards Test and a relatively low stakes writing test in New Zealand. In both cases, sample checks were employed to monitor marking.

Various unpublished works, jointly commissioned by the examination boards of England, Wales and Northern Ireland, have been used over the years to inform the debate about sample sizes for coursework moderation. Cresswell (1996), for example, compared existing examination board practices with a view to recommending a single approach and, as a result of this work, uniform procedures for the moderation of coursework were adopted. No such uniformity of practice has ever existed for written examination papers. More recently however, as part of the National Assessment Authority's quality of marking project, the debate has been augmented by Al-Bayatti & Jones (2005). They provided a discussion of the double marking sample size required to detect given levels of disagreement in marking papers. Minimum sample sizes were presented dependent upon the level of marker expertise and upon a range of detectable differences. Lamprianou (2005) also tackled this issue with a view to informing the basis of quality control measures embedded in e-marking applications.

While sampling rates have been the subject of some investigation, in none of the papers cited is there any reference to the statistically defensible derivation of the acceptable difference between two marks for a given paper or item: the marking tolerance. Perhaps if the marking tolerance is recast as the tolerable difference between two marks for a given paper, it is understandable that there is no theory behind its derivation. Indeed, some of the case studies reported in Lamprianou (2004) detailed numerical tolerance values but there was no evidence to suggest that these values were anything other than pragmatic. As Cresswell (1996) concluded:

> "technical considerations are not, of course, the only things that matter … considerations of cost and administrative ability are also very important." (p.8)

Nevertheless, the literature makes very little reference to the probability that errant markers will be identified for any given tolerable difference nor does it consider the implications of different methods of quality assurance. Existing systems seem to represent the best compromise between two conflicting imperatives - statistical robustness and practical viability - where statistical robustness is the lesser partner. By exploring two of the many methods of quality assurance, this paper considers the effect of item marking on the probability of being awarded some *true mark* and the effect of sampling on the probability of identifying an errant marker.

Two Simple Models of E-Marking and Quality Assurance

A simple model of e-marking might be that a paper is divided into its constituent parts which are then electronically distributed to markers on the basis of the level of expertise required to mark that part, or item. The marking is thus segmented and a marker will mark a number of the same items, rather than a whole paper. Items are marked on-screen with marks submitted electronically to the testing organisation.

Throughout the process marking will be checked using a quality assurance system. One such system might involve the periodic introduction of a 'seed' to a marker's allocation of work (referred to as 'validity' or 'monitor' responses by McClelland (2010)). The seed would be an item selected and pre-marked by a senior marker; often the subject expert who wrote the examination and mark scheme. Each would be selected as a clear example of a particular score and thus illustrate the application of the mark scheme. The seeds would appear to the markers as ordinary items to be scored and the mark awarded by the senior marker would not be revealed when it was presented within a marker's allocation. Failure to mark the seed, or a preset number of seeds, accurately may trigger retraining. This system is hierarchical in that the senior marker's mark would be defined as the true mark. If the marker failed to mark the seed accurately then the final mark would be that of the senior marker. Reliability, in this context, would be the level of agreement with the senior marker. Intolerable deviations from this true mark would be deemed errant.

Another system might involve the periodic double marking of a sample of the marker's allocation (referred to as 'double scoring' by McClelland (2010)). The double marking would be blind, with the original marker and mark not revealed. It could be undertaken by a peer marker and only escalated further should there be an intolerable difference between the peers. An intolerable difference, or a preset number of intolerable differences, may trigger retraining for the errant marker(s) in the pair. This system might be termed peer-pair double marking. Where the peer-pair failed, the final item mark would be that of the senior marker.

There is room for debate about the definition of true mark and about mark resolution in the case of intolerable differences between markers. In the context of this paper true mark is defined, for the hierarchical seeded quality assurance system, as the gold standard, or expert judgment, of the senior marker. Interestingly, when considering paper-based marking, Baird and Meadows (2009) showed that with large teams of markers a hierarchical structure of quality assurance promotes the establishment of communities rather than a strict adherence to the single gold standard. It is possible that these communities would not so readily exist in the more remote and mechanistic system of e-marking; particularly if marker training is held on-line with consequent reduced face-to-face contact (Chamberlain & Taylor, In Press). Nevertheless, Baird and Meadows (2009) argued that a community of practice conceptualisation of true mark should be adopted as it more realistically reflects practise. The peer-pair double marking system represents a more consensual approach but, in the simple model described, markers are still trained to follow a single centrally designed mark scheme and to make judgments independently. In resolving differences between independent judgments, the system reverts to the expert judgment of the senior marker and the reinstatement of the hierarchy.

In practice mark resolution ranges from the purely formulaic to the openly discursive. Classical test theory, for example, suggests that the true mark is derived by pooling all possible markers' marks (Spearman, 1904, 1927). Putting theory into practice, Johnson, Penny, Fisher and Kuhs (2003) summarise the literature on marking reliability when different frameworks and aggregation models are used to create a single mark for a paper. The frameworks they describe determine the structure of a quality assurance system while the aggregation models compare mechanisms to create a single mark out of many. Their discussion framework most closely follows the more democratic process favoured by Massey and Foulkes (1994); taking advantage of the full range of information available. Further as Baird and Meadows (2009) advocated, it promotes profiting from the shared experience of an examining community.

While a quality assurance system which requires discussion for every conflicting judgment is not viable in the context of national examinations, the two simple illustrative models presented herein could provide scope to maximise the influence of the community should such a change be deemed desirable. In the hierarchical seeded system, rather than the senior marker providing the true mark for a seed, this mark could be arrived at in a more consensual manner. In the peer-pair double marking system, the adjudication would not necessarily need to involve a senior marker. Furthermore, expert examining panels could be used to provide a bank of exemplar material for reference throughout marking (Baird & Meadows, 2009).

Some Simulations of Quality Assurance Systems for E-Marking

The Data & Methodology

The Data

In England, Wales and Northern Ireland, pupils at the age of 16 and in their final year of compulsory education are assessed using the national General Certificate in Secondary Education (GCSE) examinations. These examinations are offered by five examination boards.

In order to explore the effects of a hierarchical seeded or peer-pair double marking method of quality assurance, data from two GCSE assessments offered by the Assessment and Qualifications Alliance (AQA), one of the English examination boards, were used. These provided *a priori* distributions of differences between a putative true mark and the mark awarded by a particular marker. They typify the two extremes of constructed response assessment; one being a short answer paper and the other, an essay paper.

The first examination was completed by candidates in summer 2008 and formed part of a GCSE in Religious Studies. It was chosen as an example of a short answer paper with 21 items some of which were reasonably high tariff (up to eight marks). The maximum mark for the examination was 83. Data were collected from the marking process, which is currently electronic, and includes checks on quality of marking using a hierarchical seeded system.

The second examination was part of a GCSE in English and was taken in summer 2006. The paper contained two essay items each worth 27 marks. This component is not currently e-marked but was the subject of an unrelated study (Fowles, 2006). The mark-remark data collected from the study afforded the opportunity to model the effect applying both the hierarchical seeded and peer-pair double marking systems of quality assurance. The study recorded both item marks awarded by individual markers and the whole-paper mark. Therefore, the data allowed for a comparison of reliability, as defined above for the hierarchical seeded system of quality assurance, for item-marking compared with whole-paper marking.

While the Religious Studies data was collected in a live environment, it should be noted that a limitation of the English data is that it was collected under experimental conditions, after the live marking period.

Methodology

For the hierarchical seeded quality assurance system, the *a priori* distributions created from the 2008 Religious Studies examination and the 2006 English examination were used to estimate differences between the mark awarded by a junior marker and the true mark awarded by the senior marker. The estimates were made using Monte Carlo simulation. For each item within the examination, 1,000 item level mark differences were randomly generated. This created a sample of 1,000 candidates. The total question paper mark difference for each of the candidates was calculated from the item level data. The process was then replicated 100 times to allow the creation of an estimate, and associated standard error, of the difference between the mark allocated by the marker and the true mark. To contextualise the simulated data, the 100 replications might be regarded as representing 100 markers.

For the peer-pair double marked quality assurance system, the simulation followed the same principles. Rather than producing 1,000 estimates of the difference between a junior marker and the senior, it produced 500 estimates of the difference between random sets of peer-pairs. For each junior marker in the peer-pair, the same *a priori* distributions were used to estimate the difference from the senior marker. Peer-pair differences were then calculated as the difference of the differences as described in Box 1.

Let P be the senior marker's true mark

X_1 be the mark awarded by junior marker 1

X_2 be the mark awarded by junior marker 2

Now the distribution of $P - X_1$ and $P - X_2$ is described by the *a priori* distribution and can thus be used to generate random observations. From these observations an estimate of the difference between peers can be made using:

Peer difference = $(P - X_1) - (P - X_2) = X_2 - X_1$

Box 1

The simulation data were used to consider the effectiveness of the different models of quality assurance. Using data from a hierarchical seeded quality assurance system the following issues were considered:

- the effect that item marking has on the reliability of marking in papers containing short answers compared with those containing essays;
- the effect that item marking has on the reliability of marking when compared with whole script marking;
- the effect that introducing mark tolerances has on marking reliability; and
- the probability of identifying errant markers using a hierarchical seeded quality assurance system.

The probability of identifying errant markers was also considered with respect to the peer-pair double marked quality assurance system. Finally the two systems were compared in terms of defining and identifying errant markers.

A Hierarchical Seeded Quality Assurance System

Comparing Short Answer Question Papers and Essay Question Papers

The results of the simulation are shown in Figure 1. The graph shows the probability that the total mark for an examination will be less than or equal to a given absolute mark difference away from the true mark. It represents the picture that emerges from item marking if no quality assurance measures were in place and if the differences that were observed between the senior marker and the junior marker at item level were allowed to persist to the final total mark.

Almost all of the GCSE Religious Studies candidates would be awarded marks within 10% of the true mark*;* this being the point where the line approaches a probability of one. Reading from the graph, only 13% of the candidates would be awarded the true mark but nearly 80% would be within 4% of the true mark.

There is a stark contrast between the short answer Religious Studies paper and the two essay item English paper. Although the probability of being awarded the true mark is not that much lower for English (12% compared with 13% for Religious Studies), the probability of a candidate being awarded a mark within 4% of the true mark would only be about 40%.

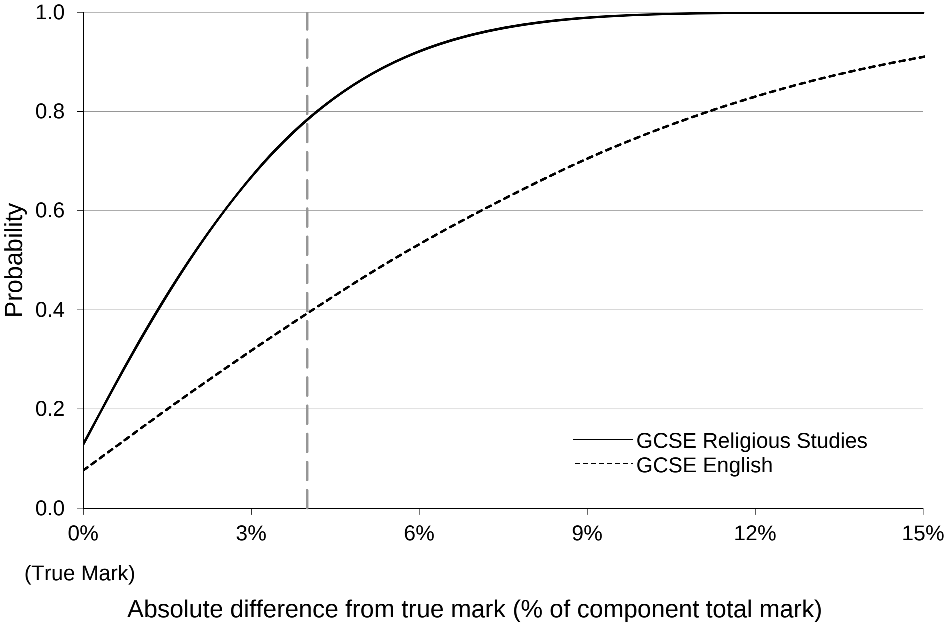

Figure 1 The probability that the total mark for an examination will be less than or equal to a given absolute mark difference away from the true mark (dotted lines denote 95% confidence intervals)

This difference in reliability, as defined in the context of the hierarchical quality assurance system, between the short answer paper and the essay paper might suggest a move towards the former format. However, the efficacy of an assessment is characterised not only by reliability (no matter how it may be defined) but also by validity. The removal of essay-type items from an English examination would clearly effect the validity of the assessment. To understand the impact on reliability of paper format is clearly imperative but this understanding must be used in conjunction with conflicting imperative of maximising validity.

Furthermore, when contextualising the data presented in Figure 1, the understanding of reliability must be operationalised by defining a tolerable difference. Clearly the level of tolerance might change over time but if, in the context of the current simulation, a 15% difference from the true mark was defined tolerable, then there would be very little conflict between the issues of reliability, as defined by the hierarchical quality assurance system, and validity. Over 90% of candidate responses for both examinations would fall within the tolerable 15% difference and therefore there would be no real advantage of one format over the other. It is worth noting, nevertheless, that the concept of reliability goes beyond that which has been defined. Amongst other things, reliability is also likely to be affected by the interaction between individual marker marking characteristics (the propensity for severity or leniency) and features of the assessment format. In the latter, the probability of an incorrect marker decision may be affected differentially by the discrete chunks of a short answer item compared with a high tariff, essay item. The untangling of these sources of error is not the subject of further discussion herein but should not be overlooked.

Comparing Item Level Marking with Whole-Paper Marking

The simulation was extended to look at the effect of item marking compared with whole-paper marking. Whole-paper marks (as produced by an individual marker) were only available for GCSE English so the analysis was limited to this examination alone. The whole-paper mark differences from the senior marker were used to create an *a priori* distribution from which to assess reliability.

The results of the simulation are shown in Figure 2. Once again, the graph shows the probability that the total mark for the examination will be less than or equal to a given absolute mark difference away from the true mark. It represents the picture that emerges from item marking if no quality assurance measures were in place and if the differences that were observed between the senior marker and the junior marker were allowed to persist to the final total mark.

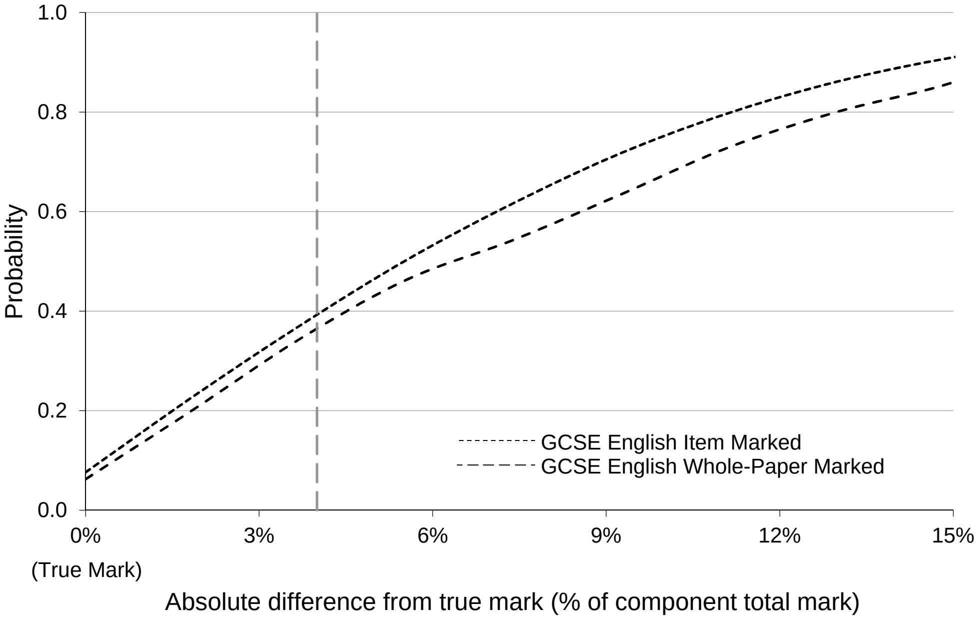

Figure 2 The probability that the total mark for the GCSE English will be less than or equal to a given absolute mark difference away from the true mark (dotted lines denote 95% confidence intervals)

The probability that the total mark for the English examination will be less than or equal to a given mark away from the true mark is greater across the range for the item marked papers. Thus the statistical theory is confirmed by empirical evidence. The use of more than one marker to mark a single paper by the introduction of item marking, offsets individual marker idiosyncrasies. The advantage of item marking is likely to be greater, the more items on a paper. Once an effective e-marking system is in place, the improvement in reliability from item marking is therefore a zero cost benefit; though it should be noted that there are cases where item marking might not be appropriate and could cause a conflict between validity and reliability. In some papers there are sequential items which are linked. The reliable marking and valid assessment of these items would require that they were seen by the same marker. Striving to increase the number of individual markers contributing to the marking of a single paper might be at the expense of introducing a valued, and valid, halo effect. So without careful partitioning of an examination, the purported benefit of e-marking in allowing random presentation of items might in fact be detrimental to the reliability of marks.

Introducing Mark Tolerances

When introduced as part of the simple model of e-marking described previously, the hierarchical quality assurance system advocated the periodic introduction of a seed to the marker's allocation. By defining a tolerable difference between the marker and his or her senior, it is feasible to estimate the best possible improvement that could be made in the reliability of marks. This estimate would be based on several assumptions:

1) reliability is defined as the level of agreement with the senior marker;
2) the senior marker's mark for an item prevails in cases where the mark difference is outside the tolerable difference; and
3) every item is double marked rather than using a seeding system.

The assumptions imply one hundred percent double marking and the infallibility of the senior marker. Operationally, complete double marking may be impossible (and pointless were the senior marker to provide the second mark) but the assumption allows an assessment of the potential for the system to minimise mark differences from the true mark.

Table 1 shows the set of mark tolerances used in the simulation to investigate the effect of hierarchical quality assurance under the assumptions specified. These tolerances are similar to those used for e-marked short answer papers in AQA.

Table 1 Mark tolerances for e-marking

Maximum Mark for the Item	Mark Tolerances
0-3	0
4-6	1
7+	2

Figure 3 shows that, in the best case scenario described above, a hierarchical quality assurance system creates a dramatic improvement in the reliability of GCSE English. This improvement, however, should be viewed cautiously as the mark tolerances were designed for short answer papers, not essay papers. Indeed the proportion of items that exceeded the mark tolerances for this assessment was over 50%; meaning that over half of the marks awarded were in fact the senior marker's mark. In this best case scenario, the hierarchical quality assurance system had very little impact on the reliability of Religious Studies examination marks.

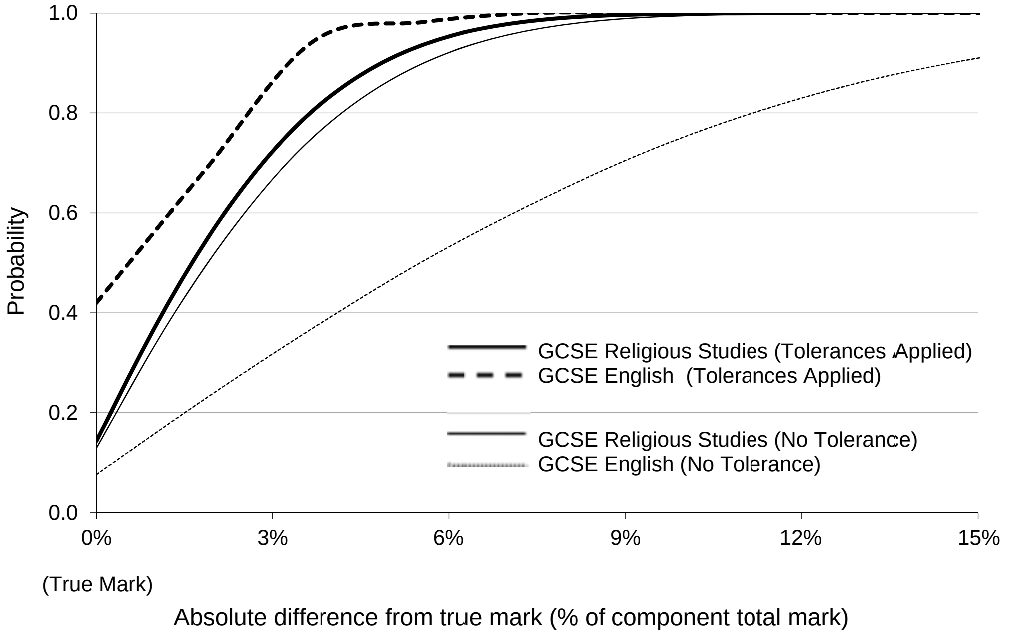

Figure 3 The probability that the total mark for an examination will be less than or equal to a given absolute mark difference away from the true mark (solid lines denote no marking tolerances and dashed lines denote tolerances set according to AQA operational rules)

Using this extreme example allows an evaluation of the use of tolerances in a quality assurance system. Too draconian and the system produces high failure rates. Too lax and the system produces little improvement in reliability. In reality, because seeds are only introduced periodically, the potential of tolerances to impact directly upon marking reliability is significantly reduced.

Suppose then that a seed failure did not result in the senior marker's mark prevailing but rather the retraining of markers. The seeds then become a method for identifying errant markers and have no direct function in improving marking reliability. The mark tolerance becomes a tolerable difference. Then the probability of markers exceeding this tolerable difference, and being identified as having done so, becomes the feature of interest.

The Probability of Identifying Errant Markers

Identifying errant markers obviously goes beyond the operational parameters set within any given e-marking and quality assurance system. With a marker hierarchy, might come an intimacy within the structure beyond that which could be gained from the perusal of statistics. Even when marking is completed remotely, markers will normally have the facility to contact others involved. This contact provides information which can not be measured and will, rightly or wrongly, give a judgmental insight into the work of a marker. Nevertheless, in a hierarchical quality assurance system, the establishment of effective operational parameters allows the addition of an objective view of marking reliability.

Using a system of e-marking, parameters such as the tolerable difference, percentage double marked and fail criteria are relatively easy to monitor and finesse. They do not represent an exhaustive list of all measures which could be said to quantify marking reliability but they are used as parameters in the simulation. For the first item on the GCSE English examination, initially the percentage of double marked seeds was varied and, subsequently, a fail criterion. Herein the fail criterion is defined as the number of items a marker marks outside the tolerable difference before that marker is deemed errant. Table 2 shows the results of a simulation where it is assumed that a marker's item allocation is 500. It should be remembered that a tighter tolerable difference implies more markers will be defined as failing. On an item marked out of 27 with a tolerable difference of two, for example, it is almost certain that an errant marker will be identified if the double marking percentage is greater than 5%. However, if the tolerable difference is set at six, and the fail criterion at three, a 10% sample will only identify just under three quarters of errant markers.

Table 2 The probability that a marker will be identified as errant given different quality assurance parameters (Hierarchical seeded quality assurance system)

Tolerable	Items > t	Percentage Double Marked			
Difference (t)	(Fail Criterion)	10%	5%	2%	1%
2	1	1.00	1.00	1.00	1.00
4		1.00	1.00	0.98	0.70
6		1.00	0.91	0.46	0.18
8		0.57	0.22	0.07	0.01
2	3	1.00	1.00	0.96	0.34
4		1.00	0.97	0.26	0.04
6		0.73	0.15	0.00	0.00
8		0.04	0.00	0.00	0.00
2	5	1.00	1.00	0.45	0.00
4		1.00	0.67	0.00	0.00
6		0.28	0.00	0.00	0.00
8		0.00	0.00	0.00	0.00

Data taken from GCSE English, item 1 (27 marks)

The standard errors associated with these figures are relatively low, all less than 0.06, but the simulation is based on a single *a priori* distribution. A limited sensitivity analysis using the second item on the GCSE English paper suggested no significant differences between the probabilities of identifying errant markers. The largest difference between the probabilities occurred in the estimate derived from a 2% double marked sample, where the fail criterion was five items marked outside the tolerable difference of two. The data for item one suggested that the probability of identifying an errant marker was 0.55 whereas for item two the probability was 0.74.

There was no systematic pattern in the differences between the probability estimates for item one and item two.

The figures reported in Table 2 correspond favourably with those reported by Al-Bayatti & Jones (2005) which were based on a 30 mark Key Stage 3 English Test. They suggested that with a double marked sample of about 5% and a tolerable difference of four, errant markers would almost certainly be identified.

The Peer-Pair Double Marked Quality Assurance System

The Probability of Identifying Errant Markers

It can be argued that a peer-pair quality assurance system might be best suited to question papers including essays because it circumvents the need for a senior marker to pre-mark seed items before widespread marking can begin. The more complex and lengthy nature of the essays means that senior marker pre-marking would introduce greater time pressures into what may be an already limited marking period. It is critical though that operational advantages are not outweighed by poorer efficacy in terms of identifying errant markers.

For the first item on the GCSE English paper, the data presented in Table 3 show that, for any given set of quality assurance parameters, the peer-pair system is more successful than the hierarchical system at identifying the markers defined to be errant. Taking the example of an item marked out of 27 with a tolerable difference of six, a fail criterion of three and a double marking sample rate of 10%, it is almost certain the errant marker will be identified using a peer-pair system. For the hierarchical system, the probability would be 0.73 (Table 2).

Table 3 The probability that an marker will be identified as errant given different quality assurance parameters (Peer-pair double marked quality assurance system)

Tolerable	Items > t	Percentage Double Marked			
Difference (t)	(Fail Criterion)	10%	5%	2%	1%
2	1	1.00	1.00	1.00	1.00
4		1.00	1.00	0.97	0.88
6		1.00	1.00	0.92	0.68
8		1.00	0.96	0.60	0.39
2	3	1.00	1.00	0.99	0.67
4		1.00	1.00	0.79	0.27
6		1.00	0.94	0.25	0.02
8		0.86	0.42	0.04	0.00
2	5	1.00	1.00	0.87	0.03
4		1.00	0.96	0.29	0.00
6		1.00	0.50	0.01	0.00
8		0.53	0.08	0.00	0.00

Data taken from GCSE English, item 1 (27 marks)

Now it could be argued that the pragmatically determined tolerable difference should be smaller for a hierarchical system than for a peer-pair system. In other words, the difference between a marker and the true mark given by his or her senior should be smaller than the difference between two peers. If a rule is established such that the tolerable difference for a peer-pair system is twice that of a hierarchical system then a comparison of Table 2 with Table 3 shows that, in fact, the hierarchical system better supports the identification of errant markers. For example, with a tolerable difference of eight for the peer-pair system and four for the hierarchical system, a 1% double marked sample and a fail criterion of one, the probability of identifying an errant marker would be 0.39 for the former and 0.70 for the latter. However, the argument for varying tolerable differences between the two models of quality assurance is only defensible if one continues to subscribe to the notion of a single gold standard.

A Comparison of the Peer-Pair Double Marked Quality Assurance System with the Hierarchical Seeded Quality Assurance System

The probability of identifying an errant maker is inextricably linked with the probability that an item mark exceeds the tolerable difference; defining the item a failure. By continuing to focus on the first item of the GCSE English paper, it is possible to compare the probability of failure using each of the quality assurance systems outlined above.

Table 4 illustrates that the probability of an item mark being deemed a failure is greater when the peer-pair quality assurance system is applied than when a hierarchy is imposed. This suggests that a single view of the true mark for an item does not exist between peers. However, the extent to which the two systems produce disparate failure rates diminishes as the tolerable difference increases; concomitantly the need to defend the notion of the single gold standard becomes less urgent even if it is regarded as the ideal.

Table 4 A comparison of item pass and fail probabilities between the hierarchical seeded and peer-pair double marked quality assurance systems

Tolerable	Peer-	Hierarchical		
Difference	Pair	Pass	Fail	Total
2	Pass	0.26	0.12	0.38
4		0.56	0.07	0.63
6		0.78	0.03	0.80
8		0.90	0.01	0.91
2	Fail	0.27	0.35	0.62
4		0.23	0.14	0.37
6		0.15	0.05	0.20
8		0.08	0.01	0.09
2	Total	0.53	0.47	
4		0.78	0.22	
6		0.93	0.07	
8		0.98	0.02	

Data taken from GCSE English, item 1 (27 marks)

In terms of defining errant markers, the peer-pair quality assurance system might be regarded as more conservative than the hierarchical system. More markers are identified as errant and therefore more remedial work (adjudication and retraining) is needed. While alleviating time pressures at the beginning of the marking period, the peer-pair system might introduce a higher volume of on-going work. Although increasing the tolerable difference between markers would decrease workload, it would be at the expense of reducing the sensitivity of the system.

Discussion & Conclusions

What emerges from the simulation exercise is that, with a system of sample double marking, the imposition of marking tolerances does little directly to improve the reliability of the mark awarded to a paper. The finding accords with Deming's (2000, p.29) assertion that "inspection does not improve quality, nor guarantee quality". Item marking, rather than whole paper marking, minimises the effect of systematic differences between markers but the extent to which this improves reliability at a paper level is dependent upon the number of items on a paper. The fewer items on the paper, the less the potential for errant marks to be compensated. Altering question paper design to increase the number of items included on a paper, however, raises serious questions about the relative importance of reliability when compared with validity. The ability to write cohesive essays is considered a key skill in many disciplines and to create an assessment which does not test this skill, but includes a large number of short answer items, would compromise validity. Furthermore, striving to increase the number of individual markers contributing to the marking of a single paper might be at the expense of allowing a valued, and valid, halo effect. So without careful partitioning of an assessment, the benefit of e-marking in allowing random presentation of items might in fact be detrimental to both the reliability and validity of marks.

In dismissing any direct effect that a quality assurance system can have on marking reliability, the system becomes a tool for identifying errant markers. An errant marker must be defined by the pragmatic imposition of parameters beyond which marking is deemed unacceptable. The key, then, is to maximise the probability of identifying the errant marker and then to take appropriate remedial action. In so doing, marking reliability can be improved indirectly by a feedback mechanism. Indeed training has been shown to have a key role in reducing marking errors even though optimal feedback mechanisms remain ambiguous. Sykes *et al*. (2009), for example, suggested that although feedback of some description was "effective in reducing marking error, … neither the type nor the amount of feedback was important in contributing to improved accuracy" (p.13). Weigle (1998) revealed that training was more effective as a tool for improving the internal consistency of marks awarded by a given marker than for arriving at a single true mark. As an aside, if mathematical compensation for individual marker characteristics is a possibility (see Baird & Mac, 1999), internal consistency might be a sufficient outcome from training. However, where peer-pair is used to identify errant markers, the removal of the single gold standard means that such mechanical recalibration is not viable. Some studies have suggested that training is best effected by a well designed mark scheme (Furneaux & Rignall, 2007; Shaw, 2002) but, no matter which feedback mechanism is implemented, it remains the case that any quality assurance system must first identify the errant markers efficiently.

In the illustration based upon GCSE English, it was shown that for any given tolerable difference, fail criterion and double marking sample rate, a peer-pair quality assurance system has a higher probability of identifying an errant marker than a hierarchical quality assurance system. The peer-pair system also has a higher probability of identifying a marking failure. The advantages of a peer-pair quality assurance system undoubtedly include the alleviation of time pressures at the beginning of the marking period and a move towards a more consensual view of the *true mark*. However, these advantages come at the expense of a single gold standard and an increased workload throughout the marking period. What becomes apparent, though, is that the differences between a hierarchical and peer-pair system of quality assurance become blurred as the parameters within the system are relaxed.

To conclude with a note of caution, although sensitivity analyses were conducted for the reported simulations, the *a priori* distributions upon which they were based were specific to particular components. There is no guarantee that the findings are representative of short answer or essay papers but they do show that reliability of marking is independent of mark tolerances. Furthermore, the simulations provide some basic tools for determining double marking sample sizes if using one of the two simple models of quality assurance. However, the simulations stop at the point of identifying errant markers. While the quality assurance models described provide scope to increase the influence of a community, more evidence is needed to confirm the value of so doing. Furthermore, questions still remain about the design of effective feedback (or retraining) mechanisms and about the extent to which the format of a question paper influences marking reliability.

References

Al-Bayatti, M. & Jones, B. (2005) *The effect of sample size on increased precision in detecting errant marking.* Report produced for the National Assessment Agency.

Baird, J. & Mac, Q. (1999) How should examiner adjustments be calculated? - a discussion paper. *AQA Internal Report,* RPA_99_JB_RC_013.

Baird, J. & Meadows, M. (2009) What is the right mark? Respecting other examiners' views in a community of practice. *IAEA 35th International Conference.* Brisbane, Australia.

Chamberlain, S. & Taylor, R. (In Press) Online or face-to-face? An experimental study of examiner training. *British Journal of Educational Technology*.

Cresswell, M. (1996) Moderation of centre-assessed components: A note concerning sample sizes. *AQA Internal Report,* RPA_96_MC_RAC_706.

Deming, W. E. (2000) *Out of the crisis,* (MIT Press).

Fowles, D. (2002) Evaluation of an e-marking pilot in GCE Chemistry: Effects on marking and examiners' views. *AQA Internal Report,* RPA_02_DEF_RC_190.

Fowles, D. (2006) How well does marking in GCSE English transfer to marking using cmi+ with annotation? *AQA Internal Report,* RPA_06_DEF_RP_047.

Furneaux, C. & Rignall, M. (Eds.) (2007) *The effect of standardisation-training on rater-judgements for the IELTS writing module,* (Cambridge, University of Cambridge ESOL Examinations and Cambridge University Press).

Hudson, G. (2009) Improving marking quality in essays – can technology help?, paper presented at the *35th IAEA Conference*, City, 12-18 September 2009.

Johnson, R. L., Penny, J., Fisher, S. & Kuhs, T. (2003) Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement In Education,* 16(4)**,** 299–322.

Lamprianou, I. (2004) *Marking quality assurance procedures: Identifying good practice internationally.* Report produced for the National Assessment Agency.

Lamprianou, I. (2005) A simulation study to inform e-marking business decisions. *AQA Internal Report,* RPA_05_BEJ_WP_013.

Lucas, A. M. (1971) Multiple marking of a matriculation biology essay question. *British Journal of Educational Psychology,* 41(1)**,** 78-84.

Massey, A. & Foulkes, J. (1994) Audit of the 1993 ks3 science national test pilot and the concept of quasi-reconciliation. *Evaluation & Research in Education,* 8(3)**,** 119-132.

McClelland, C. A. (2010) Validity response scoring: Doing it right. *R & D Connections, ETS,* 13.

Meadows, M. & Billington, L. (2005) *A review of the literature on marking reliability*. Report produced for the National Assessment Agency.

Means, B., Blando, J., Olson, K., Middleton, T., Cobb Morocco, C., Remz, R. & Zorfass, J. (1993) *Using technology to support education reform.* Report for U.S. Department of Education, Office of Educational Research and Improvement.

Pilliner, A. E. G. (1968) Examinations, in: H. J. Butcher (Ed.) *Educational research in Britain.* London, University of London Press.

Powers, D., Farnum, M., Grant, M. & Kubota, M. (1997) *A pilot test of online essay scoring.* Report for Educational Testing Service (Princeton, NJ).

Qualifications and Curriculum Authority (QCA) (2002) *Maintaining GCE A level standards:The findings of an independent panel of experts.* Report for QCA (London).

Shaw, S. (2002) The effect of standardisation training on rater judgement and inter-rater reliability for the revised CPE writing paper 2. *Research Notes,* 8.

Sheingold, K., Kane, J., Endreweit, M. & Billings, K. (1981) *Study of issues related to implementation of computer technology in schools. Final report.* Report for Bank Street College of Education.

Sheingold, K. & Tucker, M. (1990) *Restructuring for learning with technology,* (New York, L.A. Bryant, Center for Technology in Education, Bank Street College of Education).

Spearman, C. E. (1904) ‘General intelligence’ objectively determined and measured. *American Journal of Psychology,* 15, 201-293.

Spearman, C. E. (1927) *The abilities of man, their nature and measurement,* (New York, Macmillan).

Sturman, L. & Kispal, A. (2003) To e or not to e? A comparison of electronic marking and paper-based marking, paper presented at the *29th International Association for Educational Assessment Conference*, City, 5-10 October 2003.

Sykes, E., Novaković, N., Greatorex, J., Bell, J., Nádas, R. & Gill, T. (2009) How effective is fast and automated feedback to examiners in tackling the size of marking errors? *Research Matters,* 8.

Taylor, R. (2007) The impact of e-marking on enquiries after results *AQA Internal Report,* RPA_07_RT_TR_050.

Taylor, R. (2008) The impact of e-marking on enquiries after results 2006/2007 *AQA Internal Report,* RPA_08_RT_TR_019

Weigle, S. C. (1998) Using facets to model rater training effects. *Language Testing,* 15(2)**,** 263-287.

Whetton, C. & Newton, P. (2002) An evaluation of on-line marking, paper presented at the *28th International Association for Educational Assessment Conference*, City, 1-6 September 2002.

Williams, H. G. & Van Lent, L. G. (2002) *Project 2f.1: Impact of e-marking on test design.* Report for ETS Europe (Utrecht).

Wood, R. & Quinn, B. (1976) Double impression marking of English language essay and summary questions. *Educational Review,* 28(3)**,** 229-246.