36th Annual Conference of the International Association for Educational Assessment, 22-27 August 2010, Bangkok, Thailand

Presentation by Graham Hudson
Global Business Leader for Electronic Marking
DRS Data Services Limited, UK
Email: graham.hudson@drs.co.uk

Title: A quality control framework for electronic marking

ABSTRACT

The presentation will address the all-important issue of identifying poor marking in large-scale examinations and assessments. An analytical approach to reviewing the outcomes of double marking - and percentage double marking - has led to the establishment of a quality control framework which enables poor marking to be identified based upon item type, mark tariff and proportion of double-marking undertaken.

The presenter will explain how the analysis of previous examination data has enabled this framework to be developed and the use to which it can be put in identifying markers who may require additional guidance and training to keep to agreed marking standards.

The presentation will be of interest to those who manage national examinations and assessments and who wish to balance increased marking reliability with cost and practical implementation of quality control processes.

AUTHOR

Graham Hudson is Global Business Leader for Electronic Marking for DRS Data Services Limited in the UK. Graham has over twenty-five years' experience of implementing and managing large-scale assessments within the UK, including time spent at QCA conducting the marking and data collection of Key Stages 2 and 3 National Curriculum Tests. Graham now manages electronic marking for DRS for a number of awarding bodies in the UK and internationally.

1. Background

1.1 DRS has successfully implemented electronic marking with a number of awarding body clients in the UK, the largest of which is AQA. The general benefits of using electronic marking are becoming more widely recognised both within the UK and internationally.

1.2 Key to the approach adopted by DRS and its clients is the focus on improving the quality of marking through the use of technology. Marking judgements made by senior examining personnel, combined with sophisticated algorithms, enable those marking standards to be built into a marking process that continuously checks marking standards with a regularity that could not feasibly be achieved in a paper-based system.

1.3 In addition, those awarding bodies that have embarked upon exploring electronic marking have found that the change programmes initiated have led to a wider review of operational processes, leading to further streamlining and improvement that may not have been envisaged when considering electronic marking initially.

1.4 This paper provides an update to IAEA members of innovative work that has been carried out during the last year and which is now being used in a live environment.

1.5 Further detail and examples will be provided during the conference presentation.

2. Electronic marking

2.1 Electronic marking makes use of scanned images of candidates' examination and test scripts to support the marking process. Images of candidates' scripts are held securely and distributed as questions, or parts of questions, to markers for marking across the Internet. Marks are captured at the time of marking and checking of marking standards takes place in real time.

2.2 Use of the images of candidates' answers now provides many more degrees of freedom to support more rapid processing of marks and a variety of quality control measures. Paper-based systems are constrained by the physical limitations of the scripts – which can only be in one place at a time.

2.3 By dividing the candidates' scripts into segments, electronic marking provides significant improvements over conventional marking by:

- removing marking bias, related to the leniency or severity of a marker's judgement for an individual candidate and for groups of candidates;
- enabling markers to focus on topics related to their expert knowledge;
- allowing markers to focus only on marking and not be diverted by administrative or procedural matters;
- marking that does not meet the appropriate quality tolerances can be identified in real time and markers stopped from marking that item and provided with further training;
- removing clerical errors (such as addition errors by markers and transposition errors to marksheets) inherent in a paper-based system.

The most fundamental improvement, however, is enabling the regular checking of marking quality.

2.4 In addition, other processes can be supported, such as providing an electronic training resource to markers to augment or substitute the current marker standardisation meetings that take place prior to marking. This electronic process is commonly known as e-Standardisation.

3. Implementation in the UK and internationally

3.1 During the past 6 to 8 years, all UK unitary awarding bodies that provide school and further education qualifications have piloted or implemented electronic marking. A number of other, professional awarding bodies have also followed suit.

3.2 In the UK, at least 9m candidates' scripts will be scanned, imaged and marked on a PC by markers during the summer 2010. Organisations, such as DRS, have worked with awarding bodies to put in place the necessary technical infrastructure, change management, training and programme management to support the annual increase in the number of scripts processed in this way.

3.3 As a result, all major UK awarding bodies are committed to this approach and have seen the benefits identified above realised with the examiners, schools and colleges, candidates and parents.

3.4 Interest has also been expressed internationally, with DRS conducting marking pilots in Australia, the Caribbean, West Africa, Malaysia and Poland.

4. Quality control and traditional script marking

4.1 Traditional methods of quality control in general and higher education qualifications have used a mixture of approaches. The approach sometimes varies on the type of question being marked, but tends to be determined by the local environment within which the marking is being undertaken.

4.2 Essentially, two approaches can be used:

- regular sampling of work, and
- double-marking of work.

4.3 Sometimes, in other contexts, post-marking moderation of candidates' work can take place. This tends to be with smaller and more local marking panels. It can be seen as a form of sampling which may or may not lead to either further marking or some form of mark adjustment.

4.4 Of course, regular sampling is a form of double-marking, but at a defined level of intervention. Its purpose is to establish if the markers are continuing to mark at the standards set at the outset when they were trained.

4.5 Regular sampling has the following drawbacks:

- sampling is undertaken at the whole paper level, which means that systematic bias from an individual examiner can remain;
- the sample (generally) is chosen by the marker. This means that the marker could have paid especial attention to the marking of the sample papers, but not to those in between sampling;
- the number of scripts included and frequency of sampling is limited by the need to move papers between markers and supervisors (either through the post or in a marking centre);
- decisions about the acceptability of marking quality are made by supervisors who are a potential source of bias in their own right and who tend to have to make holistic decisions

on marking quality which can obscure some areas of a marker’s marking that may be inaccurate;
- poor marking quality that may remain at the end of a marking period has to be corrected – either through re-marking scripts or through statistical adjustment.

4.5 Double marking also has some drawbacks:

- setting up double-marking processes in a paper-based environment is complex and costly in its own right. Those awarding bodies internationally that have achieved this have well-thought out systems, but these are surrounded by teams of administrative staff supporting the process;
- marking at the question level is possible (and is undertaken in some places) but requires careful script management and organisation);
- as double-marking almost always takes place in a marking centre, the sampling of markers’ marking and the adjudication of difference between one marker and another, tends to take place as marking takes place. This adds stress and the risk of error because of the logistical and time constraints that exist;
- double-marking all scripts is more costly than single marking with sampling;
- as with sampling, poor marking quality that may remain at the end of a marking period has to be corrected – either through re-marking scripts or through statistical adjustment.

4.6 Both approaches require some significant investment in systems and time and ultimately do not solve the underlying need to have a regular quality checking process where intervention can take place as soon as unacceptable variances are detected.

4.7 The drawbacks are especially true with long-form answers and essays, with high mark tariffs, where markers are expected to apply professional judgement to more creative or expressive work and where variances can arise for justifiable reasons.

4.8 Electronic marking addresses all these drawbacks and enables the ‘quality plateau’ inherent in the traditional processes to be passed.

5. Quality control and electronic marking

5.1 The most common types of examination papers fall into two categories:

- candidates write their answers onto the question paper in spaces left for prose, mathematical formulae, diagrams or graphs (*constrained answer booklets*);
- candidates write their answers in free-form essay style onto a lined answer booklet without specific structure (*unconstrained answer booklets*).

5.2 Segmenting answers in a constrained answer booklet is straightforward, and all recognised electronic marking systems support this approach. Segmenting answers in an unconstrained booklet is more difficult as it is not possible to pre-determine where a candidate will begin and end an answer, although DRS has devised an approach to achieve this.

5.3 In addition, the approach to quality control will need to be different, as free-form answers tend to be longer, cover several pages and include more judgemental elements to mark. This is unlike the constrained answers which are shorter and tend to have more structured marking guidelines.

6. Quality control for 'constrained answers'

6.1 The most effective way to check marking quality for constrained answers is using 'seeded items'. This is a highly efficient way of monitoring marking standards regularly, making use of a pre-prepared bank of items marked by the senior marker team at the start of the process.

6.2 'Seeded items' are used in two ways – first at the start of each marking day to check that marking quality is correct before marking of an item is allowed; second, pairs of seeds are introduced at regular points during the marking to check that marking consistency is being maintained. Markers see the 'seeded items' as normal items to mark and will not be aware that they are quality control items.

6.3 A mark tolerance can be set that reflects the degree of agreement required between a marker's mark and the standard mark set for the 'seeded item'. For small value items, this is usually zero – in other words, the marker has to give the same mark as the standard mark. **Table 1** summarises the way in which seeded items are used.

Table 1 Summary of the use of seeded items

Type	**Detail of usage**
Qualification	A set number of seeded items are presented to a marker. Business rules are agreed with the awarding body on the number and criteria for success. For example, out of ten items presented, the agreed business rule might be that 7 out of 10 must be marked correctly to enable the marker to qualify. Other values relating to the number of qualification seeded items that can be marked differently from the seed value in a session and the maximum sum of the absolute differences between marks and seed values in a qualification session can also be set.
Marking	Pairs of seeded items are presented to the marker during the marking session. The 'gap' between the presentations of the seeded items can be set within the administration function. Two different business rules can be applied: • rule 1 – where both seeded items have to be marked correctly to continue. If one of the pair is failed, then the marker is stopped; • rule 2 – where a set number of seeds has to be marked correctly from a group of pairs marked. For example, out of the last 10 seeded items marked, 7 must be marked correctly. The parameter for setting the seed window values is expressed as a percentage, for example: • 50% gives 2 items to mark then 2 seeded items; • 20% gives 8 items to mark then 2 seeded items; • 5% gives 38 items to mark then 2 seeded items.

6.4 For all answer types, electronic marking can support various forms of double-marking. Providing images of the candidates' answers removes the traditional logistical constraints of this approach. For the more extensive free-form answers, a specific form of double-marking has been developed by DRS that makes use of the regular comparison of one marker's marking against another to keep marking within accepted tolerances. Automated or judgemental means of reconciling marking differences can be supported in real time. This is discussed further in the **Section 7** below.

6.5 The importance of segmentation and quality control methods tailored to question types cannot be underestimated, as its implementation has consequential changes in many other areas of the marking process.

7. Quality control for 'unconstrained answers'

7.1 The use of seeded items requires the establishment of a bank of items at the start of marking. This approach does not lend itself to longer answers for two reasons. One, the time taken to prepare the seeded items will be longer and two it will take markers longer to work through the seeded items before real marking can begin.

7.2 As a result, DRS has developed a set of algorithms and associated business rules that will combine the benefits of regular quality checking with those of double marking.

7.3 In so doing, a number of issues have had to be addressed, such as:

- against what standard will markers' marking be compared;
- if quality control is gauged by checking marking standards between markers, what happens to a marker when no other markers are marking;
- if mark difference exist between markers, which marker is deemed to be 'correct';
- and how does poor marking ultimately be identified and a marker stopped.

7.4 The system devised is known as *'percentage double marking'*. This means that one marker's marks are compared with another marker's marks according to a set sampling percentage. Its scope includes:

- comparing two marking opinions in 'real time';
- where differences in marking exceed a set tolerance automated business rules are used to invoke adjudication by a senior marker;
- standard items (similar to seeded items) can be used to judge (at any point in the process) how close to the 'set standard' the marking is;
- senior markers can intervene at any point to re-sample a marker's marking and, if appropriate, re-mark work for defined periods;
- combining the benefits of seeded marking and sampling marking through double marking.

7.5 There is an automated, but configurable, quality control framework in place – which uses a number of 'caps' (or limits) to manage marking quality. For a marker who 'marks ahead' of the rest, the *'pioneer cap'* comes into play and the marker is temporarily suspended from marking that particular item. This ensures that no marker can progress too far without a double check on the marking. As soon as some of the marking is marked by another marker, he or she can resume (provided no other tolerance is exceeded).

7.6 As markers mark, the number of times that a marker exceeds a set tolerance when marking is compared with other markers is recorded. When the set tolerance is exceeded, the marker is temporarily suspended from marking that item. This limit is called a *'suspect cap'*. A senior marker has to adjudicate the marking and give the 'true mark' to enable the marker to resume marking.

7.7 When the marker's mark is adjudicated and if found to be outside the tolerance of the senior marker, they accrue a *'penalty'*. There is a configurable *'penalty cap'* that will suspend a marker if too many penalties are accrued. A senior marker has to adjudicate the marking and give the 'true mark' to enable the marker to resume marking.

7.8 These mechanisms, together with the use of some pre-marked standard items, now enable long-form answers to be checked in a well-defined manner, regularly and with real-time monitoring of marking standards.

8. Building a quality control framework

8.1 A quality control framework has to have a starting point. In this case, the starting point is the question *'What is a poor marker?'*.

8.2 This is defined within the percentage double marking approach set by the e-Marker® system and will relate to the likelihood of a marker exceeding the *'penalty cap'*.

8.3 This approach has been chosen as exceeding the *'penalty cap'* is a clear status point that the markers' judgements have been examined by a senior marker and found wanting. It can be readily measured, counted and related to specific items in both terms of content domain and mark tariff.

8.4 A ***poor marker*** is defined, therefore, as '***twice as likely*** *to exceed a* ***penalty cap*** *as an* ***average marker*** *when marking a set number of average items'*.

8.5 In order to make this definition work, there is a need to set an *'optimal penalty cap'* for the average item being considered, so that a *poor marker* is likely to exceed it but an *average marker* is not. (For the purposes of this study, the optimal penalty cap was set at 1/10th of the average number of items being marked.)

8.6 In order for this framework to be of practical use, the *quantity of marking that has to be carried out for an average item to enable poor marking to be identified* has to be established.

8.10 In this case, the number of items that *optimises the difference in the probabilities of the* ***poor marker*** *and the* ***average marker*** *exceeding the penalty cap* has been chosen as the measure.

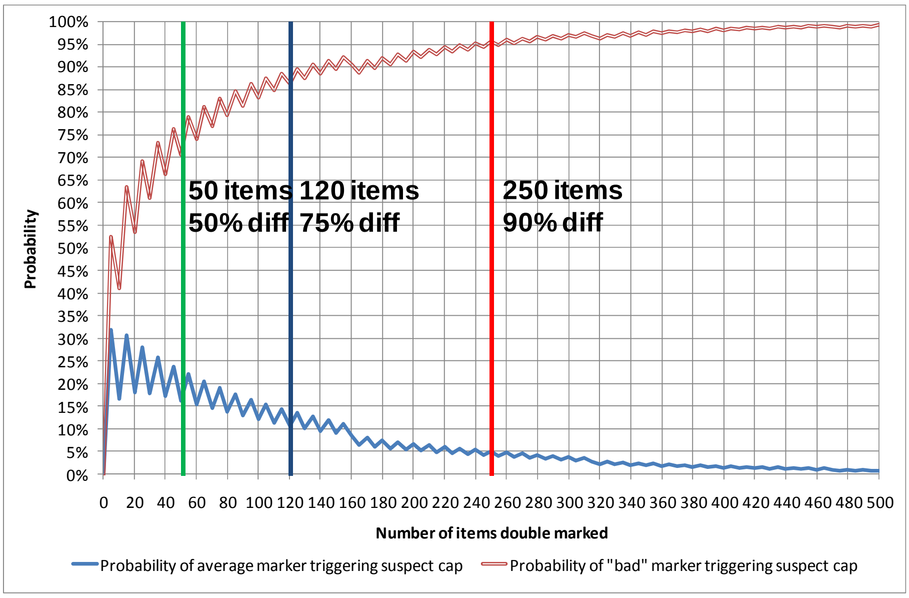
8.11 **Table 2** shows how this works out. The upper (red) line shows the probability of a poor marker exceeding the penalty cap – an increasing trend. The lower (blue) line shows the probability of an average marker exceeding the penalty cap – a decreasing trend.

8.12 The optimal difference occurs at *250 items* in this instance – where going beyond a difference of 90% in the probabilities brings little gain (shown by the red vertical line).

8.13 However, once this plot has been drawn, other lines can be reviewed that reflect other values of differences in probability and the number of items marked where detection of poor marking will occur.

8.14 So, in this example, only *120 items* would have to be marked to determine, with a *75% probability,* that a marker was poor (blue vertical line). Or, if a reduced difference in probability could be accepted, then only *50 items* would have to be marked to determine, with a *50% probability,* that a marker was poor (green vertical line).

Table 2 Difference in probability of a poor marker exceeding the optimum penalty cap and an average marker exceeding the penalty cap for an average item



8.15 It is from this point that a framework for determining how many items require marking before poor marking can be derived. This is based upon:

- the mark tariff for the item and the consequential optimal penalty cap that is determined;
- the degree of *marking risk* that an organisation is prepared to accept – ie, what probability of detecting poor marking is appropriate for the type of examination, number of candidates and use of the outcomes;
- the tolerance allowed between the mark given by the marker and the mark given by the adjudicating marker (which will change as the mark tariff increases).

8.16 Given those factors, a framework can be drawn up, as shown in **Table 3**. This shows potential different 'risk probabilities' and the number of items to mark before a poor marker should be identified. The values discussed above are shown against a 10-mark item.

Table 3 Tabulating the framework

	A poor marker – 2 times as likely to exceed the penalty cap		
	Number to double mark for difference in probability of exceeding the penalty cap greater than...		
Maximum score	**90%**	**75%**	**50%**
1	1000	490	180
2	180	90	30
:	:	:	:
9	280	140	50
10	250	120	50
:	:	:	:
15	270	130	50
16	240	120	40

8.17 This means that for an allocation of 500 average 10-mark items:

- 50% would have to be double-marked for a 90% probability of identifying the poor marker;
- 25% would have to be double-marked for a 75% probability of identifying the poor marker;
- 10% would have to be double-marked for a 50% probability of identifying the poor marker.

8.18 This provides the beginnings of an approach that should enable more certain decisions to be made concerning poor marking and how soon it can be detected. The approach to 'marking risk' will depend upon the individual organisation.

8.19 Other areas that will be explored in the future are:

1. to look to empirical ways of defining what a poor marker is (based on data taken from recent marking exercises);
2. managing the rate of double marking (eg weighting the double marking to the start of an allocation and easing off later);
3. dynamically changing the penalty cap (to find the poor marker quicker and let average markers mark).

9. Reinforcing the role of technology

9.1 The indications are that all awarding bodies that make use of electronic marking have made a point of ensuring that technology is implemented in a way that meets the needs of examinations and assessments and supports good practice. There is a risk that the use of technology can undermine important principles for the sake of, for example, logistical efficiency.

9.2 The work described here, however, has been developed in conjunction with those in the field of assessment that wish to see the reliability and accuracy of marking improved, primarily to the benefit of the candidates taking examinations.

9.3 Technology has been used to:

- bring together the best of traditional quality control mechanisms;
- put in place objective and consistent processes that are not dependent upon individual markers for their implementation;
- make the implementation of these techniques feasible.

9.4 Without technology of this kind, the ability to balance marking reliability and content validity would not be possible in high-stakes, high-volume assessment regimes. The visibility and transparency for national assessment providers that this brings is invaluable in continuing to build confidence in the outcomes for candidates.

Acknowledgement

The research analysis and framework construction described in Section 8 of this paper was carried out for DRS by the National Foundation for Educational Research (NFER), Slough, UK.