35th IAEA Conference – BRISBANE 12-18 September 2009

Presentation by Graham Hudson
Global Business Leader for Electronic Marking
DRS Data Services Limited, UK

Theme: Quality assurance and accountability in assessment

Title: Improving marking quality in essays – can technology help?

ABSTRACT

The presentation will address how marking quality in essays or long-form answers has been approached traditionally and explain how using images of candidates' answers can enable new quality mechanisms to be used.

The paper will build upon successful work already undertaken with shorter, constructed questions, which was presented at IAEA in Baku in 2007. Significant developments have been made that combine the use of scanning technology, marking algorithms and adaptable sampling criteria to provide a dynamic approach to monitoring marking variances.

The content will be relevant to those who are considering what mix of assessment formats are best used to balance marking reliability with content validity, as the processes used are focused on narrowing the gap in variance between individual markers and between markers and the agreed marking standards.

AUTHOR

Graham Hudson is Global Business Leader for Electronic Marking for DRS Data Services Limited in the UK. Graham has over 25 years' experience of implementing and managing large-scale assessments within the UK. His experience covers examination management and delivery since 1983, including time spent at the Qualifications and Curriculum Authority in the UK, conducting the marking and data collection of Key Stages 2 and 3 National Curriculum Tests.

For the past six years Graham has been managing electronic marking for DRS for a number of awarding bodies in the UK and internationally, with 20,000 markers using the tehcnology to capture over 8 million marks during 2008 in the UK alone.

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 short paper will address how marking quality in essays or long-form answers has been approached traditionally and explain how using images of candidates' answers can enable new quality mechanisms to be used.

1.5 Further detail and examples will be provided in the Seminar, *Intelligent Test Setting and Marking*, Thursday 17 September, 13.45, Ballroom 2, at the IAEA 35th Annual Conference in 2009.

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. Quality control and traditional script marking

3.1 Traditional methods of quality control in general 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.

3.2 Essentially, two approaches can be used:

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

3.3 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.

3.4 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.

3.5 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.

3.6 Even with such thorough techniques, marking quality could still be improved. Various techniques have been tried in the past, including:

- producing marking schemes that are more tightly specified;
- providing examples of candidates' scripts that illustrate particular questions and marking approaches;
- keeping marker teams small, to ensure that the supervision can be undertaken thoroughly;
- putting in place quality parameters based upon ranges of absolute mark differences across samples of marking;
- comparing like-for-like ranges of marking with objective test papers – hence identifying potential markers whose marking is systematically out of line with the cohort being marked;
- reviewing schools' distributions of marks with previous years' data to identify any statistically significant shifts that could point towards marking inaccuracy.

3.7 However, these 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.

3.8 This is 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.

3.9 Electronic marking addresses all these issues and enables the 'quality plateau' inherent in the traditional processes to be passed.

4. Quality control and electronic marking

4.1 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).

4.2 Segmenting answers in a constrained answer booklet is straightforward, and all known 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.

4.3 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.

4.4 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.

4.5 '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.

4.6 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.

Type	Detail of usage
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.

4.7 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 5** below.

4.8 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.

5. Quality control for long-form answers

5.1 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.

5.2 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.

5.3 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.

5.4 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. It involves:

- 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.

5.4 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 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).

5.5 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.

5.6 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.

5.7 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.

6. Reinforcing the role of technology

6.1 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.

6.2 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.

6.3 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.

6.4 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.

7. Next steps

7.1 As part of the marking that took place during the Summer 2009, data has been collected from a small number of subjects that made use of *percentage double marking.*

7.2 A small research exercise has been set up to evaluate the process, using researchers from the National Foundation for Educational Research (NFER) in Slough, UK.

7.3 The outcome from the study will inform further development and any enhancements to the approach that may be necessary.

7.4 A summary of the methodology to be used in provided in Annex 1 and should data be available in time, this will be presented at the IAEA Conference in Brisbane in September 2009 along with this report.

Annex 1 – Research framework

A1 The study exercise is looking at the 'closeness' of marking between markers. This has been defined by looking at every instance where an item is marked for a particular candidate and the following three variables that are being examined:

- the mark awarded by the individual marker for that particular candidate on a particular item;
- the average mark awarded across all markers who marked that candidate's response to the particular item (if the item has been marked for that student multiple times);
- the mark awarded by the senior marker for that student's response to the particular item (if the item has been checked by a senior marker for that student).

A2 Using the data, two models are being developed. One makes use of all data that has been marked multiple times and explore factors relating to the discrepancy between the marks awarded by individual markers and the average mark awarded. The second makes use of all data that has been checked by the senior marker and explore the discrepancies between marks awarded by the individual marker and the senior marker.

A3 Multilevel modelling is being used to carry out the analysis. Three different elements of discrepancy will be isolated:

- **overall discrepancy** – this is the extent to which all markers on average mark a particular response too high or too low;
- **marker bias** – this is the extent to which individual marker award marks that are consistently higher or lower than average;
- **residual variance** – this is the remaining difference not explained by the previous two elements of discrepancy.

A4 The analysis will explore the relationship between each of these elements of discrepancy and a number of other variables, including:

- **the length of time for which marking has continued** – to see if marking 'tightens up' as the period of marking continues;
- **the number of times a marker has previously been automatically identified as a concern** – which will explore the extent to which the automatic quality control mechanism is effective in improving the quality of future marking;
- **the types of interventions that have previously been used to attempt to improve marking quality with an individual marker** – which will explore the extent to which the different types of marking quality mechanisms are most strongly related to improved future quality;
- **the characteristics of the item being marked** – particularly those items made up of several sub parts and if they are more difficult or easier to mark than those made up of a single part.