

Is electronic marking just about efficiency?

Further analysis of electronic marking data to investigate factors related to marking reliability

Abstract

The paper explains some of the underpinning assessment principles of the electronic marking processes and relates these to quality criteria. In addition, data derived from examinations marked using this approach is presented with the aim of discussing how measuring examiners' marking against a population of items of known mark value can provide benefits in estimating marking reliability. Analysis of data carried out jointly by DRS and NFER is presented, based on comprehensive datasets accumulated in the process of real life electronic marking. The analysis investigates areas such as between-seed-examiner and between-seed variance, and any variations in these due to item type and other characteristics of the item, the examiner or the marking situation.

Authors

The authors of the paper are Graham Hudson, Barbara H. Donahue, Simon Rutt and Ian Schagen.

Graham Hudson is National Business Development Manager for Education for DRS in the UK. Graham has over twenty years' experience of implementing and managing large-scale assessments within the UK. His experience covers developing and managing general qualifications since 1983, including the introduction of GCSE examinations in 1988 and the National Curriculum examinations in 1994.

Graham worked at the Qualifications and Curriculum Authority for over three years during which time he ran the external marking and data collection of the Key Stages 2 and 3 tests in England and established a government-funded programme for implementing the use of new technologies in examinations and assessments.

Graham now works for DRS where he has put in place the electronic mark capture and marking of tests for a number of awarding bodies in the UK and internationally, with a total of 5.5m marks being captured using these systems in 2007.

Barbara H. Donahue received her training as a psychometrician and her PhD from the University of Georgia, Athens, Georgia, USA. While there she was involved in the standard setting process for several state-wide assessments. Since joining the NFER in 2005, she has been involved with the technical data analysis of several test development projects; year 7 and 8 optional English and Subject A and recently the key stage 2 English assessment.

Before joining the NFER in 2005, Barbara was an information analyst at Emory University (Atlanta, Georgia) responsible for data management and analysis of pharmacy, chart and Medication Event Monitoring System data as part of a 5 year multi-million dollar US National Institutes of Health funded behavioural research randomised control trial.

Simon Rutt is Deputy Head of the Statistics Research & Analysis Group at the NFER. He has been extensively involved in the production of many types of analyses and has worked closely with schools and other institutions, assisting them with engaging and understanding their data with a view to securing improvement. He is lead statistician for the Aim Higher evaluation which has looked at the factors that influence the decision to go on to higher and further education. He has worked on number of major projects for the DfES including the Fast Track to prosecution evaluation, Excellence in Cities and more recently Simon was jointly responsible for a project that looked at the achievement of ethnic minority pupils in EIC areas. Prior to joining the NFER Simon was a principal research officer at the London Borough of Hammersmith & Fulham.

Ian Schagen is Head of Statistics at the NFER with previous experience in industry and as a university lecturer. He is a Chartered Statistician and a member of the editorial board of Educational

DRS Data Services Limited National Foundation for Educational Research

Research, and is currently a member of the Research Committee of the examination board AQA. Dr Schagen has published a book, 'Statistics for School Managers' (2000), aimed at helping school staff to make use of statistical information, as well as being joint editor of a book on the use of effect sizes in educational research 'But What Does It Mean?' (2004).

Recently Ian has been involved with advising DfES about methodology for analysing the National Pupil Database (NPD), in particular as a member of the Value Added Methodology Advisory Group. He has also recently been acting as external consultant to the Department on their review of data systems underpinning their Public Service Agreement Targets. He was project director for the analysis of combined NPD/ILR data for the Learning and Skills Development Agency, looking at the impact of local patterns of post-16 provision on participation, retention and attainment, and has recently directed a project for the Learning and Skills Council to evaluate the robustness of their value-added models.

Acknowledgements

The data used for this research and which forms the basis for the paper has been provided with permission of the Assessment and Qualifications Alliance (AQA).

Summary

This paper follows on from that given by DRS at the IAEA in Baku in July 2007. That paper was based on initial findings on data collected from the Summer 2006 examinations undertaken by the Assessment and Qualifications Alliance (AQA) in England, Wales and Northern Ireland. In this paper we recapitulate on some of that work, but take it further using integrated multilevel modelling on different outcomes in order to carry out in-depth investigation of marking reliability using this data.

The reasons for undertaking the research are described and set out the framework within which the results are provided, based on the work undertaken to date. A view of marking reliability and consistency is derived from 'seed items' used to check that markers are marking to the correct 'standard'. The difference between the mark awarded for a 'seed item' by each marker and the 'standard' mark assigned to that seed has been used as the measure of marking accuracy. For the purposes of this paper, three subjects comprising nine components have been reviewed in detail.

The analysis has looked at fixed and random effects that have affected marking differences, some of which are described in the report. Linear and logistical regression and cross-classified multi-level modelling have been used. Areas for further investigation have also been noted.

A high very degree of agreement was noted between markers' marking and the standard seed item marks. The residual error has been modelled in detail using factors related to seed items, examiners and the circumstances under which marking was carried out, with some findings of interest. The main differences between this paper and the version presented at Baku are:

- the fixed and random parts of the analysis have been integrated into a single set of multilevel models;
- three different outcomes have been considered: algebraic award difference, absolute award difference, and probability of passing the seed;
- three subjects have been considered, with a total of nine papers;
- information on item type has been included in the modelling.

The report concludes that the marking accuracy is very high overall and provides valuable information on how to improve the business rules that determine how the overall quality control model is run operationally.

1. Background

AQA and DRS have worked together successfully to introduce electronic marking to an increasing number of GCE and GCSE examinations in England, Wales and Northern Ireland. During 2007, 105 examination components were marked from scanned images, with over 1.7 million candidates' scripts being processed. Over 2,000 examiners accessed the marking system from their homes via the internet and marked 43 million items. A further 15 million items were marked by senior examiners from candidate answers that had been captured electronically. Using other mark capture applications, a further 2.7 million candidates' marks were collected electronically from the original paper scripts.

The efficiency benefits of using electronic marking have been rehearsed in the past by AQA and DRS and have been the subject of previous papers to the IAEA. However, both AQA and DRS see major gains in relation to improved marking accuracy as being vital to bringing improvements to the examining system in the UK. The current suite of applications being used to carry out this work is described in **Annex 1**.

Central to the management of marking quality and consistency is the use of 'seed items'. Unlike conventional methods of checking marking quality with paper scripts, which rely on periodic sampling, the use of 'seed items' enables marking quality to be checked at an item level as marking takes place. Markers who do not mark to the correct standard can either be retrained on an item or stopped from marking that item altogether.

2. The use of 'seed items'

'Seed 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.

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 'seed 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 2.1** summarises the way in which seeds are used.

Table 2.1 Summary of the use of seeds

| Type | Detail of usage |
|---------------|---|
| Qualification | <p>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, 7 out of 10 must be marked correctly to enable the marker to qualify.</p> <p>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.</p> |

**DRS Data Services Limited
National Foundation for Educational Research**

| Type | Detail of usage |
|---------|---|
| Marking | <p>Pairs of seeded items are presented to the marker during the marking session. The 'gap' between the presentation of the seeded items can be set within the administration function. Two different business rules can be applied:</p> <ul style="list-style-type: none"> • 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. <p>The parameters for setting the seed window values are expressed as a percentage, for example:</p> <ul style="list-style-type: none"> • 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. |

3. The scope of the study

Thirteen examination components from the AQA Summer 2006 examinations were chosen for the study. All the data was examined and the detailed work was focused on three subjects from different disciplines, which have been identified as Subjects A to C. This was done to enable comparisons to be made between Subjects that had different item types and to control the data volume to be examined at this stage. **Table 3.1** shows the details of the information available at the start of the study.

Table 3.1 Data available at the start of the study

| Component | Number of Candidates | Number of Markers | Number of Parts | Number of Seed Examiners | Number of Seeds | Number of Seed Events |
|-------------------|----------------------|-------------------|-----------------|--------------------------|-----------------|-----------------------|
| Subject A Paper 1 | 23,716 | 51 | 51 | 8 | 2,055 | 53,847 |
| Subject A Paper 2 | 23,716 | 60 | 56 | 9 | 1,716 | 61,153 |
| Subject B Paper 1 | 25,343 | 51 | 41 | 7 | 1,763 | 71,208 |
| Subject B Paper 2 | 22,131 | 98 | 37 | 7 | 1,681 | 194,880 |
| Subject B Paper 3 | 70,270 | 44 | 37 | 6 | 1,552 | 70,007 |
| Subject C Paper 1 | 9,009 | 30 | 54 | 2 | 813 | 16,633 |
| Subject C Paper 2 | 14,200 | 36 | 34 | 3 | 1,118 | 37,357 |
| Subject C Paper 3 | 10,870 | 33 | 37 | 3 | 1,100 | 25,353 |
| Subject C Paper 4 | 11,660 | 32 | 34 | 3 | 1,021 | 27,929 |
| Subject D Paper 1 | 15,383 | 38 | 34 | 6 | 1,429 | 51,719 |
| Subject E Paper 1 | 134,060 | 221 | 46 | 31 | 3,496 | 400,688 |
| Subject E Paper 2 | 134,060 | 247 | 44 | 19 | 2,406 | 390,645 |
| Subject F Paper 1 | 52,248 | 72 | 34 | 2 | 479 | 33,077 |
| Total | 546,666 | 1,013 | 539 | 106 | 20,629 | 1,434,496 |

DRS Data Services Limited National Foundation for Educational Research

The key to the data columns is as follows:

| | |
|---------------------------|--|
| Number of candidates: | Number of candidates whose total marks were captured on the database |
| Number of markers: | Number of markers involved in marking the candidates' papers |
| Number of parts: | The number of discrete items to be marked on the paper |
| Number of seed examiners: | The number of senior examiners involved in setting the standard mark for each seed used |
| Number of seeds: | The total number of seeds for all items that had been created for use by the system |
| Number of seed events: | The total number of times all seeds had been used by the markers marking the items in each paper |

Diagrams 1 and 2 illustrate the types of items set in each of the papers and illustrate the differences in the type and length of response expected from the candidates in similar Subjects.

Diagram 1 – Types of questions set in Subject A Paper 2
Reproduced with the permission of AQA

(iii) Since 1970, the government of the Maldives has made rules that have to be followed when building any new tourist development.
The table below lists some of these rules.

| | |
|---|---|
| 1 | Resorts are to use recycled water in the gardens. |
| 2 | No buildings are to be taller than the tree-tops. |
| 3 | No more than 20% of any island is to be built on. |
| 4 | Each island is to have its own solar-powered generator for producing electricity. |

Choose **three** of these rules, and suggest why each was felt to be important.

Rule number

.....

.....

.....

Rule number

.....

.....

.....

Rule number

.....

.....

.....

(6 marks)

DRS Data Services Limited National Foundation for Educational Research

Diagram 2 – Types of questions set in Subject B Paper 1
Reproduced with the permission of AQA

4 (a) A sequence of numbers is shown.

3 7 11 15

Write down the next two numbers in the sequence. (2 marks)

(b) Another sequence of numbers is shown.

3 7 12 18

Write down the next number in this sequence. (1 mark)

(c) A different sequence begins

3 6 12 24 48

Write down a rule for this sequence.

Answer

.....
(1 mark)

The quantity of data available for review was considerable and provided a wealth of opportunity for reviewing marking comparisons at the item level that would not be available from conventional marking approaches.

To narrow the work, nine components were chosen to undertake the specific awarding difference studies – two Subject A , three Subject B and four Subject C. **Table 3.2** shows a summary of the awarding differences seen for these components taken from the total of seed events.

Table 3.2 The total number of award differences for selected components

| Award Difference | Subj A Paper 1 | Subj A Paper 2 | Subj B Paper 1 | Subj B Paper 2 | Subj B Paper 3 | Subj C Paper 1 | Subj C Paper 2 | Subj C Paper 3 | Subj C Paper 4 |
|------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| -5 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| -4 | 2 | 31 | 0 | 3 | 3 | 1 | 1 | 3 | 0 |
| -3 | 30 | 217 | 6 | 22 | 14 | 7 | 17 | 7 | 4 |
| -2 | 372 | 1256 | 83 | 93 | 184 | 16 | 185 | 21 | 73 |
| -1 | 3264 | 5561 | 481 | 1570 | 841 | 478 | 1674 | 692 | 1036 |
| 0 | 46782 | 48318 | 70300 | 191818 | 67646 | 15728 | 33927 | 24274 | 25825 |
| 1 | 2977 | 4933 | 329 | 1216 | 1128 | 371 | 1414 | 345 | 868 |
| 2 | 375 | 736 | 6 | 112 | 168 | 30 | 132 | 8 | 111 |
| 3 | 44 | 84 | 3 | 45 | 14 | 2 | 7 | 1 | 12 |
| 4 | 1 | 12 | 0 | 1 | 9 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Exact agreement | 86.9% | 79.0% | 98.7% | 98.4% | 96.6% | 94.6% | 90.8% | 95.7% | 92.5% |

Award differences of zero indicate that the marker and the seed examiner were in exact agreement. For the nine examination papers, the exact agreement was very high, from 79.0% to 98.7%. Overall, there was very little variability in award difference.

A tolerance value can be set that will allow the examiner to mark the seeded item acceptably, but not have to agree exactly with the seed value. The value of the tolerance, if greater than zero, will depend upon the nature of the question and its total mark value.

It should be noted that some of the award differences will include tolerance values that are greater than zero (with a greater proportion of these being in the Subject B components). Once exact agreement has been accounted for, 45% of the remaining seed events were within tolerance. Therefore, this data alone should not be seen as the measure of marking accuracy. This will be discussed further in later sections.

Preliminary analysis of the seed marking data, reported in the earlier paper, considered separately the fixed effects of background factors and the random effects of different seeds, different markers and different items. It also focused on just two subjects (A and B above) with five papers. However, some interesting results were produced, which are summarised below:

The 'fixed effects' analysis showed the following:

- qualification seeds were less likely to be deemed acceptable;
- in both Subject A papers and one Subject B paper (Paper 1), seeds marked more often tended to be more acceptable – in the other two Subject B papers (Paper2 and Paper 3) the opposite was true;
- by and large, the more often a part was encountered the more likely the seed was to be marked acceptably – the exception was Subject B Paper 3;
- 9-5 Markers for Subject B were more likely to mark seeds acceptably;
- in general slightly higher seed marks were given for seeds deemed to be acceptable;
- in two Subject B papers, 9-5 markers tended to mark seeds very slightly higher.

From Section 5, on the 'random effects' analysis:

- error variances attributed to markers and items are minimal;
- the variance attributed to seeds is approximately one-third of the total;
- the residual 'noise' variance accounts for approximately two-thirds of the total;
- the overall variance is lower for Subject B than it is for Subject A;
- the average award difference is not significantly different from zero for three out of five components – it is significantly negative for Subject A paper 2 and Subject B paper 1.

In the following section we shall describe further analysis integrating the fixed and random analysis into a unified multilevel model, with a number of outcomes relating to marking reliability, controlling for a range of background factors and assuming a hierarchy of seed examiners, the seeds they produce, and the marking events involving those seeds.

4. Further Multilevel Analysis of Seed Marking Data

In the previous paper we considered two approaches to analysing the seed marking data separately: a random effects model to consider the different components of the total mark uncertainty; and a fixed effects model to look for characteristics of the markers, the items or the marking which could be related to mark variability. We also used a single measure of mark variability as an outcome, and only considered two subjects (Subject A and Subject B) with a total of five different papers. In the current study we have extended our analysis in several ways:

- taking account of information on the type of item being marked;
- analysing three different measures of mark variability instead of just one;
- extending the number of subjects and papers considered by including Subject C, with four additional papers;
- combining the random and fixed elements of the model into a single multilevel model for analysis, taking account of the hierarchical nature of the data.

Item types

Information on item types was matched into the general dataset comprising all the seed marking events being considered. Basically there were two item types: those which could be marked by a 'general' marker who was not a subject expert (labelled 'G'), and those which had to be marked by a subject expert (labelled 'E'). It is important to realise that these labels applied to the items not to the

DRS Data Services Limited National Foundation for Educational Research

markers – an expert marker could mark a ‘G’ item while a general marker should not have tackled an ‘E’ item. However, by identifying as an expert marker anyone who marked one or more ‘E’ items it was also possible to characterise markers, and this information is included in **Table 4.1** below.

Table 4.1 Item and Marker Types by Subject

| Subject | Items | | | Markers | | |
|-------------------|-------|---------|-------|---------|--------------|------------|
| | Total | No. ‘E’ | % ‘E’ | Total | No. ‘expert’ | % ‘expert’ |
| Subject A Paper 1 | 37 | 31 | 84% | 44 | 23 | 52% |
| Subject A Paper 2 | 42 | 25 | 60% | 51 | 39 | 77% |
| Subject A Paper 3 | 47 | 34 | 72% | 60 | 51 | 85% |
| Subject B Paper 1 | 40 | 23 | 58% | 51 | 29 | 57% |
| Subject B Paper 2 | 37 | 29 | 78% | 98 | 94 | 96% |
| Subject C Paper 1 | 27 | 8 | 30% | 30 | 13 | 43% |
| Subject C Paper 2 | 31 | 15 | 48% | 36 | 21 | 58% |
| Subject C Paper 3 | 32 | 10 | 31% | 33 | 17 | 52% |
| Subject C Paper 4 | 31 | 11 | 35% | 32 | 18 | 56% |

Outcome measures

Three different measures of seed marking variability were used in this analysis:

1. The difference between the mark awarded by the examiner on the seed and that originally awarded by the seed examiner, as an algebraic difference. Positive values imply lenience on the part of the examiner while negative values imply severity. In general, this is a measure of bias in the marking by examiners compared with seed examiners.
2. The absolute value of the above difference, as a measure of the gap between the examiner’s mark and that awarded by the seed examiner. In general, this is a measure of marking accuracy.
3. Whether or not the absolute difference defined above was ‘acceptable’, in that it fell within the predefined tolerances for this seed. This is another measure of marking accuracy, transformed to a binary (0/1) indicator.

Setting up multilevel models

For this particular analysis, the multilevel models were set up with three levels:

1. The seed examiner who developed the seed;
2. The seed itself;
3. The ‘marking event’ when an examiner marked the seed.

In principle, there may be random effects due to the actual items on which the seeds are based, but some analysis using cross-classified multilevel models (see Goldstein, 2003) demonstrated that this added little or nothing to the explanatory power of the model so a simple hierarchical structure was retained. The explanatory variables included in the fixed part of the model were as follows:

- whether or not the seed was a ‘qualification seed’;
- number of times the examiner had previously seen the same seed;
- the volume of marking carried out by the examiner on this paper: broken into quintiles, with the first (lowest) quintile indicator taken as the default;
- at what time of day the marking was done: morning (9.00 to 11.59), afternoon (12.00 to 16.59), evening (17.00 to 20.59) or late (21.00 to 0.59); with the default being early (5.00 to 8.59);
- whether the marker only marked during office hours (morning and afternoon);
- the maximum mark available on the item;
- whether the item was an ‘expert’ item or not.

Analysis was carried out using the multilevel modelling package MIWiN (See Rasbah et al, 2000), with all background variables initially included in the model for each of the three outcomes for each of the 9 papers. Background variables which were non-significant were progressively removed from each model until a parsimonious fit to the data was achieved with only those factors statistically significant at the 5% level left. In the case of the third outcome (probability of passing the seed) logistic multilevel analysis was used in order to model the binary outcome. Results of this analysis are presented in the following section.

Results of multilevel modelling

In order to present the results of modelling, it is not particularly helpful to show the coefficients of all significant factors, as the size of these may depend on the scale of the background variable concerned as well as the strength of the relationship with the outcome. For this reason, results are presented in terms of 'quasi effect sizes' (see Schagen, 2004, for more details) which measure the strength of the relationship between each variable and the outcome in dimensionless terms. Roughly speaking, they represent the change in the outcome, as a percentage of its standard deviation, associated with an 'average change' in the background variable. **Tables 4.2, 4.3 and 4.4** show the significant background variables for each subject and paper in terms of their 'quasi effect sizes', for each of the three outcomes described earlier.

The random part of each model is also of interest, as it quantifies the total 'noise' in the seed marking process and partitions it between seed examiners, seeds, and the seed marking events. **Tables 4.5, 4.6, and 4.7** summarise the random variances modelled for each paper, focusing only on the first two outcomes – the random variances for the logistic outcome are more difficult to interpret. We show the variances at each level, the total variance and corresponding standard deviation in the award differences (both algebraic and absolute), and the percentages of the variance at each level.

Interpretation of results

One of the main purposes of this analysis was to find variables which were associated with mark variability in order to see how the latter could be reduced by modifying relevant factors. The first thing to be said is that the amount of mark variability to be explained is quite low in this data. Examination of **Tables 4.5, 4.6 and 4.7** shows that the total standard deviation in mark variability is a fraction of a mark for all papers, demonstrating that there is relatively little scope for further reduction. In general, the overall variance is divided such that two-thirds is related to the seed marking event, one-third to the seed itself, and virtually nothing to the seed examiner. The implication of this might be that tightening up on the quality of seeds might reduce total variation by up to one-third.

Having said this, **Tables 4.2, 4.3 and 4.4** show the relationships between measures of mark variability and the various background factors. These are considered below, for each background factor in turn.

- **Qualification seed:** By and large, qualification seeds tend to have higher absolute award difference and are less likely to be passed. In Subject A, markers tend to be slightly more lenient for qualification seeds, and there is a mixed picture for Subject C and no significant relationship for Subject B in terms of severity/leniency.
- **Number of times seen seed:** In two of the Subject B papers, the more a seed has been seen the lower the absolute award difference and the more likely it is to be passed. This is also true for one Subject A paper and two of the Subject C papers. The opposite is true for the other Subject A paper. The picture in terms of algebraic difference (severity/leniency) is mixed, but only three papers in total have any significant effect.
- **Total seeds marked:** There are overall quite mixed results for this, with no clear pattern. In some cases (e.g. Subject B Paper 1, Subject A Paper 2, some Subject C papers) the more marking is done, the higher the award difference and the lower the probability of passing. In other cases (e.g. Subject A Paper 1 in particular) the opposite is very much the case.

DRS Data Services Limited
National Foundation for Educational Research

- **Time of day:** Again, results are fairly variable, with few significant coefficients and little clear pattern. In a few cases, marking late at night is associated with higher absolute differences and lower probability of passing.
- **Working office hours only:** This variable is related to the markers, and has a fairly close relationship to the variable on item type, in that those marking 'general' items tend to work office hours and those marking 'expert' items tend to mark at other times as well. This relationship is not perfect in most cases, so it is possible to disentangle the two effects when both are considered together. For all the Subject C papers, and Subject B Paper 3, markers working office hours only tended to have higher probabilities of passing the seed as well as (to some extent) lower absolute award differences. For Subject A Paper 2, however, the opposite was the case.
- **Maximum marks available:** The general pattern is that items with more marks available tend to have higher absolute award differences, and for Subject B and Subject A this is also associated with lower probabilities of passing the seed. For two Subject C papers, however, the latter effect is reversed – possibly implying that in these cases the seed tolerances have been adjusted to compensate for the wider mark ranges.
- **'Expert' items:** In Subject B Paper 1 the 'expert' items tend to be associated with lower absolute award differences, whereas in other cases (Subject A Paper 1 and three Subject C papers) the reverse is true. However, in only one Subject C paper is there a negative relationship between 'expert' items and the probability of passing a seed. It seems that there is probably a complex relationship between item type, marker expertise, marking variability and seed tolerances.

DRS Data Services Limited
National Foundation for Educational Research

Table 4.2 Quasi Effect Sizes for Significant variables related to Subject A Paper Outcomes

| Variable | Paper 1 | Paper 2 |
|--|---------|---------|
| Outcome: Algebraic mark difference (Examiner – seed examiner) | | |
| Qualification seed | 1.8 | 2.5 |
| Number of times seen seed | -4.8 | 2.7 |
| Total marked – 2nd quintile | | |
| Total marked - 3rd quintile | 7.9 | -2.6 |
| Total marked - 4th quintile | 10.6 | |
| Total marked - 5th quintile | 11.7 | -7.1 |
| Marked in morning | | |
| Marked in afternoon | -2.3 | |
| Marked in evening | -4.3 | |
| Marked late at night | | |
| Works office hours only | | |
| Max mark available on item | | |
| 'Expert' item | | |
| Outcome: Absolute award difference | | |
| Qualification seed | 4.0 | 2.4 |
| Number of times seen seed | 2.7 | -6.7 |
| Total marked - 2nd quintile | -4.8 | -4.0 |
| Total marked - 3rd quintile | -5.2 | |
| Total marked - 4th quintile | | 5.0 |
| Total marked - 5th quintile | -4.3 | 8.5 |
| Marked in morning | | |
| Marked in afternoon | | |
| Marked in evening | | |
| Marked late at night | | |
| Works office hours only | | 14.3 |
| Max mark available on item | 33.3 | 50.4 |
| 'Expert' item | 34.4 | |
| Outcome: Probability of passing seed | | |
| Qualification seed | -10.0 | -10 |
| Number of times seen seed | -16.5 | 31 |
| Total marked - 2nd quintile | 15.2 | 9 |
| Total marked - 3rd quintile | 15.1 | |
| Total marked - 4th quintile | 18.3 | -14 |
| Total marked - 5th quintile | 71.7 | -27 |
| Marked in morning | | |
| Marked in afternoon | | |
| Marked in evening | | |
| Marked late at night | | |
| Works office hours only | | -53 |
| Max mark available on item | -78.5 | -15 |
| 'Expert' item | | |

**DRS Data Services Limited
National Foundation for Educational Research**

Table 4.3 Quasi Effect Sizes for Significant variables related to Subject B Paper Outcomes

| Variable | Paper 1 | Paper 2 | Paper 3 |
|--|---------|---------|---------|
| Outcome: Algebraic mark difference (Examiner – seed examiner) | | | |
| Qualification seed | | | |
| Number of times seen seed | | | |
| Total marked - 2nd quintile | | | |
| Total marked - 3rd quintile | | | |
| Total marked - 4th quintile | | 1.9 | 6.0 |
| Total marked - 5th quintile | | | 2.6 |
| Marked in morning | | | |
| Marked in afternoon | -2.6 | | |
| Marked in evening | -2.1 | 1.9 | |
| Marked late at night | | | |
| Works office hours only | | | |
| Max mark available on item | | | 9.8 |
| 'Expert' item | | | |
| Outcome: Absolute award difference | | | |
| Qualification seed | 2.1 | 2.2 | 2.0 |
| Number of times seen seed | -3.1 | -2.1 | |
| Total marked - 2nd quintile | | 1.4 | |
| Total marked - 3rd quintile | | | |
| Total marked - 4th quintile | | | |
| Total marked - 5th quintile | | | |
| Marked in morning | | | |
| Marked in afternoon | | | |
| Marked in evening | | | |
| Marked late at night | | | 1.8 |
| Works office hours only | | | |
| Max mark available on item | 6.3 | 15.7 | 25.8 |
| 'Expert' item | -16.9 | | |
| Outcome: Probability of passing seed | | | |
| Qualification seed | -5.8 | -8.7 | -6.9 |
| Number of times seen seed | 19.6 | 8.4 | |
| Total marked – 2nd quintile | -5.9 | -5.0 | |
| Total marked – 3rd quintile | -10.0 | -3.9 | |
| Total marked – 4th quintile | -14.4 | | |
| Total marked – 5th quintile | | | |
| Marked in morning | | | |
| Marked in afternoon | | | |
| Marked in evening | | | |
| Marked late at night | | | -5.6 |
| Works office hours only | | | 46.4 |
| Max mark available on item | | -9.9 | -9.5 |
| 'Expert' item | | | |

**DRS Data Services Limited
National Foundation for Educational Research**

Table 4.4 Quasi Effect Sizes for Significant variables related to Subject C Paper Outcomes

| Variable | Paper 1 | Paper 2 | Paper 3 | Paper 4 |
|--|---------|---------|---------|---------|
| Outcome: Algebraic mark difference (Examiner – seed examiner) | | | | |
| Qualification seed | -6.3 | 2.0 | | 3.3 |
| Number of times seen seed | | 3.8 | | |
| Total marked - 2nd quintile | | | | 8.6 |
| Total marked - 3rd quintile | -8.2 | | | -8.2 |
| Total marked - 4th quintile | | | | |
| Total marked - 5th quintile | | | | |
| Marked in morning | | | | |
| Marked in afternoon | | | | |
| Marked in evening | | | | |
| Marked late at night | | | | -5.2 |
| Works office hours only | -6.7 | | | |
| Max mark available on item | 11.8 | | -4.7 | 8.6 |
| 'Expert' item | | | | |
| Outcome: Absolute award difference | | | | |
| Qualification seed | 8.1 | | 5.1 | 6.0 |
| Number of times seen seed | -7.0 | | | |
| Total marked - 2nd quintile | | -2.4 | | |
| Total marked - 3rd quintile | 8.0 | | | |
| Total marked - 4th quintile | 16.4 | | | |
| Total marked - 5th quintile | 11.6 | | | |
| Marked in morning | 5.3 | -4.3 | | |
| Marked in afternoon | 7.4 | -4.2 | | |
| Marked in evening | | -4.1 | | |
| Marked late at night | 8.8 | | 8.3 | |
| Works office hours only | | -39.2 | -6.5 | -15.5 |
| Max mark available on item | 33.0 | 36.0 | 22.4 | 34.5 |
| 'Expert' item | 42.7 | | 11.6 | 27.6 |
| Outcome: Probability of passing seed | | | | |
| Qualification seed | -14.4 | -9.0 | -11.2 | -16.1 |
| Number of times seen seed | 17.6 | | | 31.2 |
| Total marked – 2nd quintile | | | -9.2 | |
| Total marked – 3rd quintile | -16.0 | | | -19.0 |
| Total marked – 4th quintile | -22.5 | -9.2 | | -32.3 |
| Total marked – 5th quintile | | | | |
| Marked in morning | | | | |
| Marked in afternoon | | | | |
| Marked in evening | | | | |
| Marked late at night | | | | |
| Works office hours only | 22.5 | 62.7 | 45.2 | 68.2 |
| Max mark available on item | -13.2 | | 24.7 | 39.6 |
| 'Expert' item | -34.4 | | | |

**DRS Data Services Limited
National Foundation for Educational Research**

Table 4.6 Random variances for Subject A Papers

| Variance | Paper 1 | Paper 2 |
|--|---------|---------|
| Outcome: Algebraic mark difference (Examiner – seed examiner) | | |
| Seeder variance | 0.00000 | 0.00073 |
| Seed variance | 0.05289 | 0.13840 |
| Event variance | 0.12470 | 0.22380 |
| Total variance | 0.17759 | 0.36293 |
| Standard deviation in mark | 0.421 | 0.602 |
| Seeder variance (% of total) | 0.0% | 0.2% |
| Seed variance (% of total) | 29.8% | 38.1% |
| Event variance (% of total) | 70.2% | 61.7% |
| Outcome: Absolute award difference | | |
| Seeder variance | 0.00187 | 0.00273 |
| Seed variance | 0.03299 | 0.06763 |
| Event variance | 0.11010 | 0.18430 |
| Total variance | 0.14496 | 0.25466 |
| Standard deviation in mark | 0.381 | 0.505 |
| Seeder variance (% of total) | 1.3% | 1.1% |
| Seed variance (% of total) | 22.8% | 26.6% |
| Event variance (% of total) | 76.0% | 72.4% |

Table 4.7 Random variances for Subject B Papers

| Variance | Paper 1 | Paper 2 | Paper 3 |
|--|---------|---------|---------|
| Outcome: Algebraic mark difference (Examiner – seed examiner) | | | |
| Seeder variance | 0.00000 | 0.00000 | 0.00006 |
| Seed variance | 0.00611 | 0.00852 | 0.02217 |
| Event variance | 0.01216 | 0.01715 | 0.03385 |
| Total variance | 0.01827 | 0.02567 | 0.05608 |
| Standard deviation in mark | 0.135 | 0.160 | 0.237 |
| Seeder variance (% of total) | 0.0% | 0.0% | 0.1% |
| Seed variance (% of total) | 33.4% | 33.2% | 39.5% |
| Event variance (% of total) | 66.6% | 66.8% | 60.4% |
| Outcome: Absolute award difference | | | |
| Seeder variance | 0.00015 | 0.00001 | 0.00026 |
| Seed variance | 0.00595 | 0.00921 | 0.02020 |
| Event variance | 0.01200 | 0.01615 | 0.03244 |
| Total variance | 0.01811 | 0.02537 | 0.05290 |
| Standard deviation in mark | 0.135 | 0.159 | 0.230 |
| Seeder variance (% of total) | 0.9% | 0.0% | 0.5% |
| Seed variance (% of total) | 32.9% | 36.3% | 38.2% |
| Event variance (% of total) | 66.3% | 63.7% | 61.3% |

**DRS Data Services Limited
National Foundation for Educational Research**

Table 4.8 Random variances for Subject C Papers

| Variance | Paper 1 | Paper 2 | Paper 3 | Paper 4 |
|--|---------|---------|---------|---------|
| Outcome: Algebraic mark difference (Examiner – seed examiner) | | | | |
| Seeder variance | 0.00038 | 0.00042 | 0.00018 | 0.00000 |
| Seed variance | 0.02954 | 0.04273 | 0.01392 | 0.03132 |
| Event variance | 0.04441 | 0.07883 | 0.03985 | 0.06677 |
| Total variance | 0.07433 | 0.12198 | 0.05395 | 0.09809 |
| Standard deviation in mark | 0.273 | 0.349 | 0.232 | 0.313 |
| Seeder variance (% of total) | 0.5% | 0.3% | 0.3% | 0.0% |
| Seed variance (% of total) | 39.7% | 35.0% | 25.8% | 31.9% |
| Event variance (% of total) | 59.7% | 64.6% | 73.9% | 68.1% |
| Outcome: Absolute award difference | | | | |
| Seeder variance | 0.00000 | 0.00038 | 0.00021 | 0.00057 |
| Seed variance | 0.02230 | 0.03129 | 0.01298 | 0.02246 |
| Event variance | 0.04217 | 0.06866 | 0.03705 | 0.05887 |
| Total variance | 0.06447 | 0.10033 | 0.05024 | 0.08190 |
| Standard deviation in mark | 0.254 | 0.317 | 0.224 | 0.286 |
| Seeder variance (% of total) | 0.0% | 0.4% | 0.4% | 0.7% |
| Seed variance (% of total) | 34.6% | 31.2% | 25.8% | 27.4% |
| Event variance (% of total) | 65.4% | 68.4% | 73.7% | 71.9% |

5. Conclusions to date

The electronic marking system allows for the collection of vast amounts of data on marker accuracy. This, in turn, provides the opportunity for examining factors that might affect marker accuracy and to hypothesise how to minimise their effect, as there will always be a small amount of random error (background noise) in the system.

From this preliminary work on analysing the rich and complex data available on marker accuracy from the seeding system, we can already identify some tentative conclusions based on the analysis of three different subjects. Specifically, the results of this exercise show that there is very little variability in the system. The exact agreement is very high and as a result what is left is very little variation in award difference. This provides considerable confidence for awarding bodies using the system that the investment is being seen not only in operational efficiency but, more importantly, in marking accuracy.

General findings from this analysis include:

- with slight variations between papers, about one-third of the unexplained variance in mark variability was explained by differences between seeds, and most of the remaining variance was random 'noise' at the marking event level. The variance due to seed examiners was minimal, less than 1% in most cases;
- there were no consistent background variables associated with algebraic award difference, i.e. the severity/leniency of the marking relative to seed examiners;
- absolute award difference, i.e. marking variability, tended to be higher for qualification seeds and seed items with larger numbers of marks available. In a few papers there was evidence that it was higher for seeds which were marked late at night;
- the probability of passing an item had similar relationships with background factors to those for absolute award difference, except in the opposite sense: higher absolute differences were related to lower probabilities of passing the seed, and vice versa;

DRS Data Services Limited National Foundation for Educational Research

- there were complex interactions between item type ('expert' or 'general'), papers, whether markers only marked in office hours, and mark variability. There were some interesting patterns within certain papers and subjects, but nothing which operated consistently across Subjects;
- overall mark variability was small, with standard deviations in absolute award difference being no more than a fraction of a mark.

The analysis of the random effects showed that seed variance accounted for about 30% of the variability in award difference. This analysis allowed the separation of this variability from the random error or noise in the system. Future work could be undertaken to examine how seeds are developed in order to explore ways to decrease seed variance. In particular, perhaps a system could be examined whereby multiple seed examiners agreed to (a) a particular part from a particular script being appropriate for seeding and (b) to the number of marks to be awarded for that particular script/part, before it could be put forward in the seeding pool.

The power of electronic marking is that not only does it allow us to quantify the accuracy of the marking system, but also to collect detailed information which can be analysed in such a way as to provide clues for improving reliability even further. The work outlined in this paper is a first step towards this goal.

References

Goldstein, H. (2003). *Multilevel Statistical Models*. 3rd edn. London: Arnold.#

Rasbah, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. and Lewis, T. (2000). *A user's guide to MLwiN*. Version 2.1a edn. London: Institute of Education.

Schagen, I. (2004). 'Presenting the results of complex models - normalised coefficients, star wars plots and other ideas.' In: Schagen, I. and Elliot, K. (Eds) *But what does it mean? The use of effect sizes in educational research*. Slough: NFER.

DRS Data Services Limited National Foundation for Educational Research

ANNEX 1

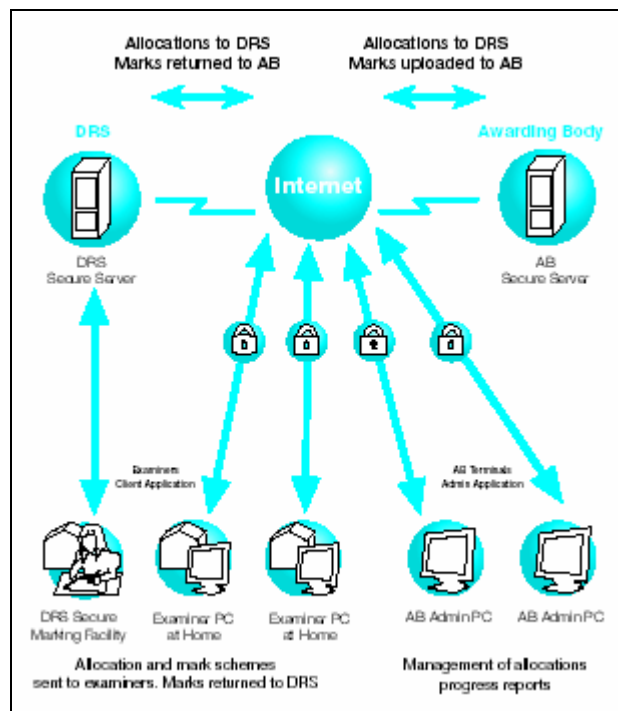
Description of e-Marker® applications

Capabilities

The applications have been designed to fit with awarding bodies needs – whatever the number of examinations or candidates are being marked. The internet suite of applications has been extended for 2006 and can be summarised as below:

| | |
|--|---|
| On-Screen Marksheets (OMS) | <i>Allows the input of total component marks direct onto screen, replacing paper-based mark capture forms</i> |
| Question Marking from Script (QMS) | <i>Allows the input of item marks direct onto screen, once scripts have been marked</i> |
| Computer Marking from Multimedia (CMM) | <i>Similar to QMS, but allows the input of marks from audio tapes for speaking components</i> |
| Computer Marking from Script (CMS) | <i>Allows the direct marking of scripts onto screen, capturing item marks directly</i> |
| Computer Marking from Image (CMI) | <i>Allows the direct marking of images of complete scripts onto screen, capturing item marks directly.</i> |
| Computer Marking from Image+ (CMI+) | <i>Allows the direct marking of individual items directed to specific markers determined by marking capability and item type.</i> |

An overview of the current system is shown in the following diagram:



DRS Data Services Limited National Foundation for Educational Research

Benefits for markers and awarding bodies

A summary of benefits of all applications mentioned is shown in the table below. The major benefits realised in 2005 relate to the detailed management information that can be derived from the CMI⁺ application. The item level data provides information for awarding bodies that was available previously. A change to the way that the quality of marking is judged has also provided much closer control over marking standards in real time, as well as providing a more detailed analysis of marking quality.

| Benefits | OMS | QMS | CMS | CMI | CMI ⁺ |
|--|-----|-----|-----|-----|------------------|
| Real-time marking management | ■ | ■ | ■ | ■ | ■ |
| Identify anomalies and missing scripts earlier | ■ | ■ | ■ | ■ | ■ |
| Regular performance monitoring | | ■ | ■ | ■ | ■ |
| No postage delays returning scripts to the awarding body | | | | ■ | ■ |
| Faster transfer of marks | ■ | ■ | ■ | ■ | ■ |
| Auto totalling of marks | | ■ | ■ | ■ | ■ |
| No answers can be missed | | ■ | ■ | ■ | ■ |
| Mark parameters handled | | ■ | ■ | ■ | ■ |
| Centralised mark schemes | | | ■ | ■ | ■ |
| Full image of script available | | | | ■ | ■ |
| e-Sampling and seeding capabilities | | | | ■ | ■ |
| No paper script sent to markers | | | | ■ | ■ |
| Electronic re-allocation of scripts and items | | | | ■ | ■ |
| Improved support for grade awarding | | | | ■ | ■ |
| Item specialisation | | | | | ■ |
| Less call on expert marking | | | | | ■ |
| Automatic marking | | | | | ■ |
| Increased general marking | | | | | ■ |
| Escalation of marking problems to an adjudicator | | | | | ■ |

A key benefit that underpins the business case for electronic marking is the ability to differentiate item marking by type and marking approach. This allows for the differentiation of the cost of marking as well as providing more information on the marking process.

DRS Data Services Limited National Foundation for Educational Research

The use of the administration application provided to awarding bodies provides access to detailed operating and quality information that leads to other benefits, as follows:

Script-based and CMI components*

- set up of component parameters, marker types and rank and administrators;
- tracking of marking by total marks;
- tracking of sampling;
- matching of unexpected candidates with entry details.
- exporting of completed marks.

CMI components only*

- tracking of marking by item;
- direct management of marking quality through seeding;
- image viewing for awarding and other purposes.