33rd IAEA Conference – BAKU 17-21 September 2007
Presentation by Graham Hudson, National Business Development Manager, Education, DRS Data Services Limited, UK

Is electronic marking just about efficiency?

Abstract

The presentation will explain some of the underpinning assessment principles of the electronic marking processes and relate these to quality criteria.

The theme of the paper will concern the use of electronic capture and marking of candidates' answers to examination scripts to improve marking reliability. The answer-types discussed will be mainly short answer and short paragraph, with some extended answers.

Little research has been carried out on the reliability of marking in this context and the presentation will be led by specially commissioned research evidence gained from data derived from examinations marked using scanning and imaging of answers. Data derived from measuring markers' marking against a population of items of known mark value will be examined to establish what benefits can be gained in marking reliability.

Analysis of data carried out would be presented, based on comprehensive datasets accumulated in the process of real-life electronic marking. The analysis would look at areas such as between-marker and between-script variance, and any variations in these due to item type, for example. Areas of future development will also be presented.

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 mathematics 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 BA, MSc, 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 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

The last presentation that DRS gave at the IAEA in Singapore looked at validity, reliability and bias in the electronic marking arena. This paper takes the area of reliability and consistency in marking further and is 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.

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 report, two subjects only 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 very high degree of agreement was noted between markers’ marking and the standard seed item marks. The factors identified so far that affect the degree of agreement include the subject being marked, the amount of marking a marker has done, the nature of the items and the ‘seed items’ used. The residual error noted will be the subject of further work.

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 to successfully introduce electronic marking to an increasing number of GCE and GCSE examinations in the 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	A set number of seeded items is 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 to mark any further items that day. 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 presentation 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 parameters for setting the seed window values are 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.

3. The scope of the study

Thirteen examination components from the AQA Summer 2006 examinations were chosen for the study. All data were examined and the detailed work was narrowed down to two subjects from different disciplines and which have been identified as Subject A and Subject B. 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	15,383	38	34	6	1,429	51,719
Subject D Paper 1	134,060	221	46	31	3,496	400,688
Subject D Paper 2	134,060	247	44	19	2,406	390,645
Subject E Tier F	9,009	30	54	2	813	16,633
Subject E Tier H	14,200	36	34	3	1,118	37,357
Subject F Tier F	10,870	33	37	3	1,100	25,353
Subject F Tier H	11,660	32	34	3	1,021	27,929
Subject G Tier F	52,248	72	34	2	479	33,077
Total	**546,666**	**1,013**	**539**	**106**	**20,629**	**1,434,496**

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

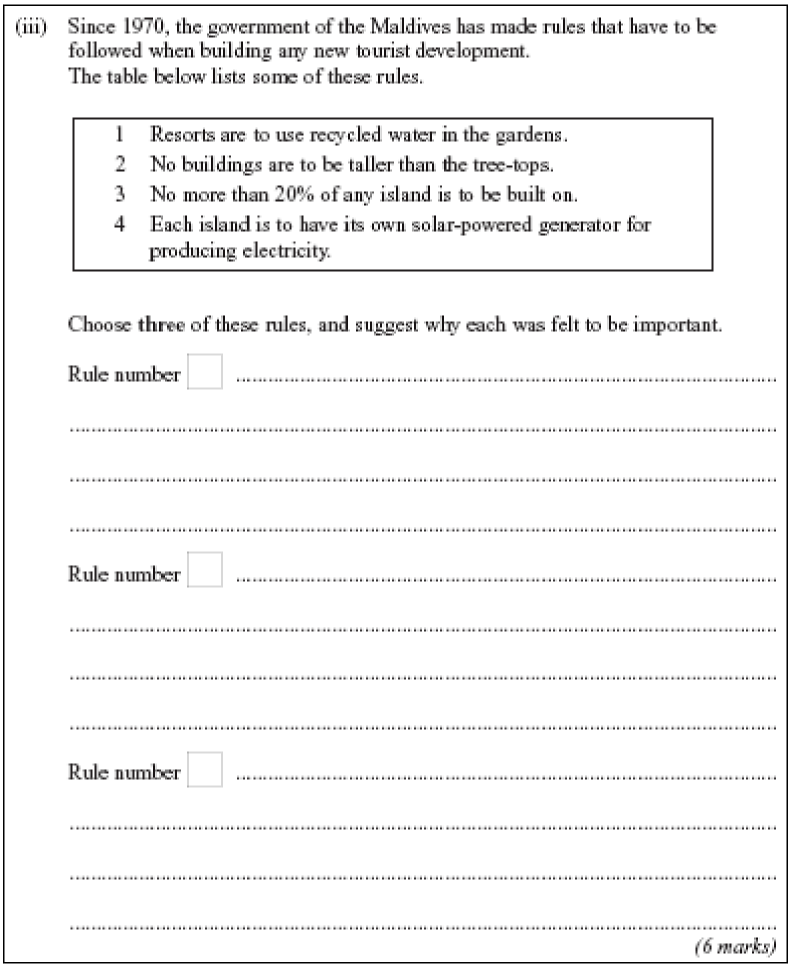
Number of seeds: The total number of seeds for all items that had been created for use by the system

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 □ ....

....

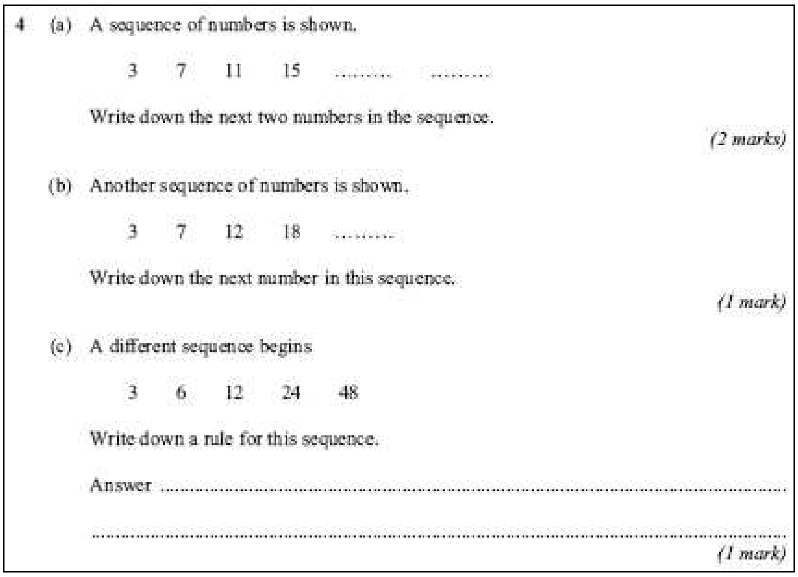
....

....

(6 marks)

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, five components were chosen to undertake the specific awarding difference studies – two Subject A and three Subject B. **Table 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 Differenc e	**Subject A Paper 1**	**Subject A Paper 2**	**Subject B Paper 1**	**Subject B Paper 2**	**Subject B Paper 3**
-5	0	5	0	0	0
-4	2	31	0	3	3
-3	30	217	6	22	14
-2	372	1256	83	93	184
-1	3264	5561	481	1570	841
0	46782	48318	70300	191818	67646
1	2977	4933	329	1216	1128
2	375	736	6	112	168
3	44	84	3	45	14
4	1	12	0	1	9
5	0	0	0	0	0
Exact agreement	86.9%	79.0%	98.7%	98.4%	96.6%

Award differences of zero indicate that the marker and the seed examiner were in exact agreement. For the five 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 A 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.

Two approaches to the analysis have been undertaken. One was to examine factors affecting award difference by component. The other was to examine factors affecting award difference by markers, items and seeds. In addition, a cross-tabulation Kappa analysis of award differences was carried out to get an initial view of marking consistency at a component level. The next two sections of the paper will provide more details of the two approaches.

Initial findings indicate the following:

- the more marking experience that the markers gained, the less likely they were to mark the ‘seed items’ incorrectly;
- some markers marked the same seed more than once and marked it differently than on previous occasions;
- the effect of item parts on mark award differences is negligible;
- the marker effect on mark award differences is slightly greater than the item effect, but is also negligible;
- the effect of seeds on mark award differences accounts for about 30% of the total variance of mark differences, but is still small when the overall effect is taken into account;
- the remaining ‘background noise’ effects account for about 60% of the total variance of mark differences, but again these were small when the overall effect is taken into account;
- the overall degree of marking accuracy as defined by mark award differences is very high – especially in Subject B.

4. Fixed effect analysis

To understand the relationship between the marking of seeds and the available explanatory variables two different models were run. The first outcome identified was whether the seed part was acceptable; to try and understand this outcome measure logistic regression models were run.

Logistic regression is a form of regression analysis in which the outcome of interest is binary, i.e. just takes two values – for example: a seed being acceptable or not being acceptable. A set of background variables can be used to predict the probabilities of the binary outcome, as in conventional regression analysis, but the coefficients relate to increasing or decreasing the probability that an outcome occurs.

Logistic regression deals with the relative *odds* associated with an event, which are equal to:

$$\frac{\text{Probability of event occurring}}{\text{Probability of event not occurring}}$$

The procedure gives an *odds ratio,* which compares the odds of an event (e.g. a seed being acceptable) associated with one group of markers, with the odds for another group. An odds ratio close to one indicates there is little difference between two groups, whereas an odds ratio significantly greater or less than one indicates differences in seed marking between the groups.

The second outcome identified was the difference in seed mark between the examiner and the seed examiner. This was a variable with a score ranging from -5 to 5 and therefore allowed a linear model between explanatory variables and outcome to be fitted.

There were a number of explanatory variables that could have been put into the models but it was decided to concentrate on the amount of marking an examiner carries out on the same part, when they do the marking, how much marking they do of the same seed and whether the seed was a qualification seed or not.

The amount of marking carried out was determined by counting the number of times an examiner saw the same part within a subject paper. This, for each subject paper, was then split into quintiles so that an examiner who saw the most parts was in the highest quintile. This allowed us to look at any relationship between the amount of marking and an examiner's likelihood of agreeing with the seed examiner. A count of the number of times an examiner saw the same seed was also created, and this enabled us to look at whether seeing the same seed increased the likelihood of agreeing with the seed examiner.

Examiners, within the constraints of the system, are able to organise their marking sessions. Some examiners mark in the morning, some late in the evening and some mark throughout the day. To understand whether this decision affects the likelihood of agreeing with the seed examiner, two variables were created that identified the highest and lowest proportion of work carried out in a single session. A single session was identified as being either early in the day, in the morning, in the afternoon, in the evening or late in the day.

There are two types of seeds within the data: seeds for qualifying and non-qualifying seeds. Qualification seeds are presented at the beginning of the marking day while non-qualifying seeds are presented throughout the marking process. If markers were going to be inaccurate it is best to happen at the beginning of the day, rather than during marking. By adding a variable that identified whether a seed was a qualification seed or not we were able to identify any differences in outcome that may be explained by this.

An additional variable was included for the Subject B models that identified those examiners who marked only between 9am and 5pm during the day. These are identified as ‘9-5 Markers’.

[It should be noted that very straightforward items are sent to ‘General Markers’ for marking whilst those that require domain knowledge are sent to ‘Expert Markers’. ‘General Markers’ tend to work office hours and do not work in quite the same way as ‘Expert Markers’. Some of the analysis below will comment upon the marking of seeded items by ‘9-5 markers’ and some of the effect may be attributable to the type of item being marked. This will be an area for further investigation.]

Results

Regression models were run for each of the two outcomes identified above (seed mark acceptable, and mark difference from ‘seeder’) and run separately for the two Subject A papers and the three Subject B papers. For these models seeds that had been retired have been excluded from the analysis. (Seeds can be removed from the system if required – which is known as ‘retiring’ a seed.)

As already identified from the descriptive analysis the level of agreement between seed examiner and examiner in terms of overall seed acceptance was extremely high and the difference in actual award was extremely small. This therefore means the models are trying to explain a very small amount of actual variance in outcome and as there is a general amount of noise, which will be explored in the next chapter, the models do not have very much explanatory power. Even so there are some significant effects and some subject differences. For reporting purposes a significance level of 0.01 was used. Coefficients which are significant at this level are shown in bold in the tables that follow. **Table 4.1** summarises the results of the logistic modelling for each of the five components studied.

Table 4.1 Logistic regression results – odds ratios for each component

Variable	**Subject A Paper 1**	**Subject A paper 2**	**Subject B Paper 1**	**Subject B Paper 2**	**Subject B Paper 3**
9-5 Markers			**1.994**	**3.392**	**4.663**
Lowest marking rate	**1.012**	1.001	1.000	**0.985**	1.000
Highest marking rate	**1.018**	0.997	0.997	**0.990**	0.995
2nd quintile part mark rate	0.925	1.014	0.901	0.864	0.946
3rd quintile part mark rate	**0.644**	1.052	0.965	0.917	0.781
4th quintile part mark rate	**0.592**	0.992	1.174	1.169	**0.693**
5th quintile part mark rate	**4.659**	1.015	1.930	**1.574**	0.947
Number of times seen seed	**1.113**	**1.108**	1.032	0.981	0.978
Qualification seed	**0.733**	**0.836**	**0.780**	**0.717**	**0.783**

For all papers an examiner was less likely to have an acceptable seed if that seed was a qualification seed. For Subject A papers, they were more likely to have an acceptable seed the more they saw the same seed. For Subject B papers, those who marked between 9am and 5pm were more like to have an acceptable seed. There were few other consistent relationships uncovered by this analysis, except that for two papers (Subject A paper 1 and Subject B Paper 2) those in the highest quintile of marking rate were more likely to have a seed mark accepted.

For the linear regression models, with award difference between examiner and seed examiner as the outcome, an additional variable was included that identified whether the seed was acceptable. As it was possible to have a difference in award but to still have an acceptable seed, this variable would control for this effect. In **Table 4.2** below, only coefficients which are significant at a level of 0.05 are included, due to the 'step forward' procedure used to fit the models[1]. Standardised coefficients are shown, which are equivalent to partial correlations with the outcome measure (award difference) controlling for other factors. Those which are significant at the 0.01 level are shown in bold.

Table 4.2 Linear regression results – standardised coefficients for each component

Variable	**Subject A Paper 1**	**Subject A paper 2**	**Subject B Paper 1**	**Subject B Paper 2**	**Subject B Paper 3**
Seed is acceptable	**0.035**	**0.167**	**0.288**	**0.012**	**0.032**
9-5 Markers				**0.022**	**0.026**
Lowest marking rate	**-0.015**		0.008		
Highest marking rate		**-0.012**			
2nd quintile part mark rate	-0.010	**0.013**			
3rd quintile part mark rate	**0.010**	**-0.012**			
4th quintile part mark rate				0.005	0.010
5th quintile part mark rate		**-0.033**			
Number of times seen seed		**0.029**		-0.013	-0.012
Qualification seed		**0.020**			

In all components there was a tendency for acceptable seeds to be marked higher than by the seed examiner. For two Subject B papers (Paper 2 and Paper 3) this was also the case for those who marked between 9am and 5pm. There were some other significant relationships on individual components, but none which were consistent across components.

These logistic and linear regression models should be regarded as preliminary attempts to look at the available data in order to see which factors may be associated with marker reliability and bias. Further work in this area can clearly be done – for example, we might want to look at absolute award difference as a measure of unreliability, and to investigate in more detail the relationships between marking rates, time of marking, and so forth. In addition, the combination of this 'fixed effect' analysis with the 'random effects' analysis described in the next section would provide a powerful way forward.

[1] The 'step forward' algorithm automatically adds new variables to a regression model until no more are significant at a given level (e.g. 5%). See Marriott, F. (1990), p.197.

To summarise briefly the results of this analysis:

- qualification seeds were less likely to be deemed acceptable. (This indicates that the quality mechanism is fulfilling its purpose here as markers become reacclimatised to the marking standard at the beginning of a marking session.);
- 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 (Paper 2 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.

5. Award difference and random effect analysis

In the previous section we focused on what are technically known as fixed effects – factors which consistently tend to increase or decrease award differences. In this section we shall focus on random effects – those aspects of the variation in award difference which cannot be directly explained, but which may be associated with elements of the underlying structure of the data, such as seeds, items or markers. To some extent we are trying to investigate the underlying ‘noise’ and attribute different parts of it to different parts of the system.

Kappa coefficients

One way of investigating the amount of inconsistency between markers and seed examiners is to use a measure of disagreement between them. Cohen (1980) developed a measure of exact agreement between raters which allows for the probability of chance agreement – this measure, Cohen’s kappa, ranges from zero (purely chance agreement) to one (exact agreement in all cases). It is intuitively obvious that the chance of exact agreement is likely to be reduced when there is a wider range of marks available. For this reason, calculations of kappa for this data have been broken down according to the maximum marks available for the seed item. Results have also been calculated separately for each of the five different components (two for Subject A and three for Subject B). **Table 5.1** shows the kappa coefficients based on all marking events broken down by subject component and maximum item marks, and the same information is presented graphically in **Figure 5.1**.

Table 5.1 Cohen's Kappa Values for Marker/Seeder Agreement by Subject Component and Maximum Item Marks

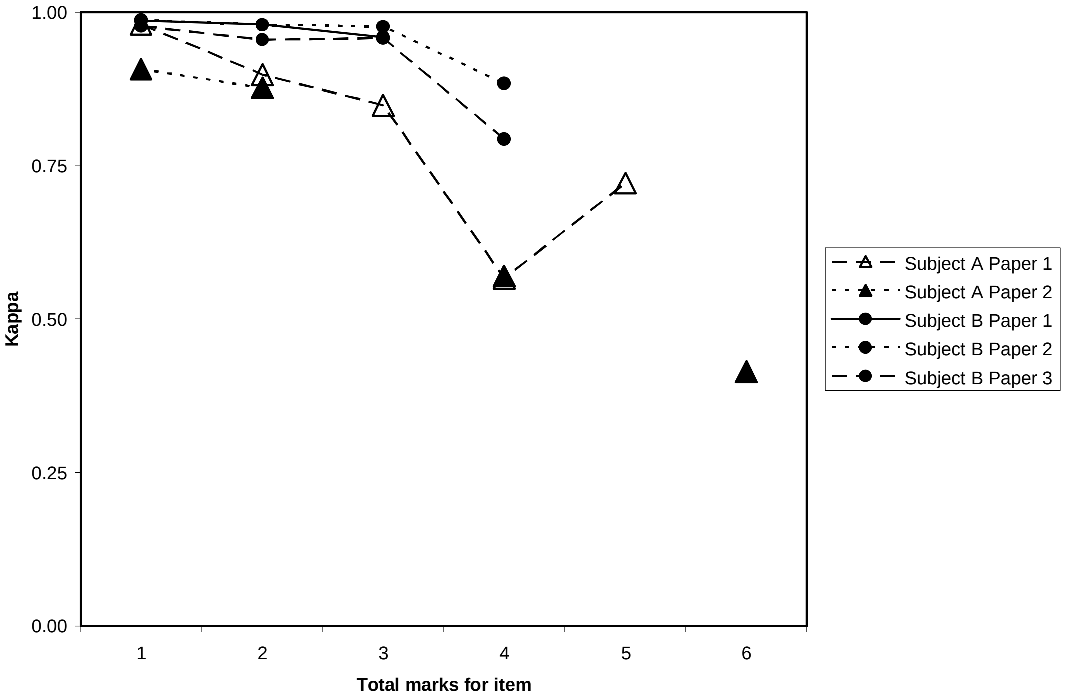
Number of marks	Subject A Paper 1	Subject A Paper 2	Subject B Paper 1	Subject B Paper 2	Subject B Paper 3
1	0.9801	0.9073	0.9863	0.9875	0.9772
2	0.8984	0.8766	0.9797	0.9798	0.9552
3	0.8485		0.9593	0.9767	0.9576
4	0.5662	0.5706		0.8839	0.7934
5	0.7220				
6		0.4143			

From these results, there are two fairly clear indications:

- Kappa values tend to decline with maximum item marks, as expected;
- Kappa values are generally lower for Subject A than for Subject B.

The converse of kappa is the amount of disagreement, or unexplained variation, in the marker data. We shall now proceed to investigate this in some detail for the same five subject components separately.

Figure 5.1: Cohen's Kappa Values for Marker/Seeder Agreement



Partitioning the variance

The overall variance in award differences may be attributed to different sources:

1. Differences between markers – some may be overall ‘lenient’ and others ‘severe’;
2. Differences between items selected – some may lead to ‘lenient’ and other to ‘severe’ marking;
3. Differences between seeds selected – as above;
4. Residual error or ‘noise’ – for example, markers giving different marks when they see the same seed.

In order to build a model which allows us to investigate the relative importance of these different sources of error we need to consider its structure. We assume a model of the following form:

$$y_{(i)jkl} = \mu + q_i + s_{(i)j} + r_k + \varepsilon_{(i)jkl} \qquad (5.1)$$

where:

$y_{(i)jkl}$ is the award difference for the *j*th seed related to the *i*th item for the *k*th marker at the *l*th marking event;

μ is the overall average award difference;

q_i is the effect of the *i*th item;

$s_{(i)j}$ is the effect of the *j*th seed related to the *i*th item;

r_k is the effect of the *k*th marker;

$\varepsilon_{(i)jkl}$ is the residual error for the *j*th seed related to the *i*th item for the *k*th marker at the *l*th marking event;

From the above it is clear that the model is not a simple hierarchical (multilevel) one, but has markers crossed with seeds and items. To analyse data in this way requires a ‘cross-classified’ multilevel model, which can be fitted using the software MlwiN (see Rasbash et al, 2000, pp254ff). Five such models have been run, one per subject component, and the results are summarised in **Table 5.2** below.

Table 5.2 Cross-Classified Multilevel Model results by Subject Component

	Subject A Paper 1	Subject A Paper 2	Subject B Paper 1	Subject B Paper 2	Subject B Paper 3
Number of markers	51	60	51	98	44
Number of items	42	47	40	37	37
Number of seeds	2055	1716	1763	1681	1552
Total cases	53847	61153	71208	194880	70007
Percentages of variation					
Marker variance	1.6%	1.2%	0.2%	0.1%	0.1%
Item variance	*0.0%*	0.9%	*0.3%*	*0.5%*	1.2%
Seed variance	29.8%	37.4%	33.1%	32.7%	38.6%
Residual variance	68.6%	60.5%	66.4%	66.8%	60.1%
Overall variance and average award difference					
Total variance	0.1782	0.3635	0.0183	0.0257	0.0563
Standard error	0.42	0.60	0.14	0.16	0.24
Overall average award difference	*0.0001*	-0.0351	-0.0061	*0.0019*	*0.0026*

(Values in italics are not statistically significant at the 5% level)

From the above results we may draw the following conclusions, for these subject components at least:

- 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.

Although we ran these models as cross-classified, the results show that the effects due to individual markers being biased are very small, so this element could be omitted from future modelling.

6. Conclusions to date and next steps

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 two 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.

From Section 4, on the ‘fixed effects’ analysis:

- qualification seeds were less likely to be deemed acceptable. (This indicates that the quality mechanism is fulfilling its purpose here as markers become reacclimatised to the marking standard at the beginning of a marking session.);
- 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.

It is clear that this kind of data modelling can provide some insights, but also that we need to go further in order to gain a deeper understanding of the complex relationships involved, and hence to improve marker accuracy even further.

Next steps in modelling the data currently held will include:

- continuing to work at the component (subject paper) level, rather than trying to create a single over-arching model for all subjects as this now looks to be the most fruitful area to explore;
- discounting item and marker effects from the random part of the model, removing the need for cross-classified models;
- including seed examiner as a level in the multilevel model;
- including ‘seed item’ effects – seeders, markers, items, seeds, marking time, seed creation date and time to attempt to explain this variation further.

The analysis of the random effects showed that seed variance accounted for nearly 30% of the variability in award difference. This analysis allows the separation of seed 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 overall reliability of the final mark even further. The work outlined in this paper is a first step towards this goal.

References

Cohen, J. (1980) ‘A coefficient of agreement for nominal scales’, in *Educational and Psychological Measurement, 20***,** pp.37-46.

Marriott, F.H.C. (1990) *A Dictionary of Statistical Terms, 5th edition.* Singapore: Longman.

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.

ANNEX 1

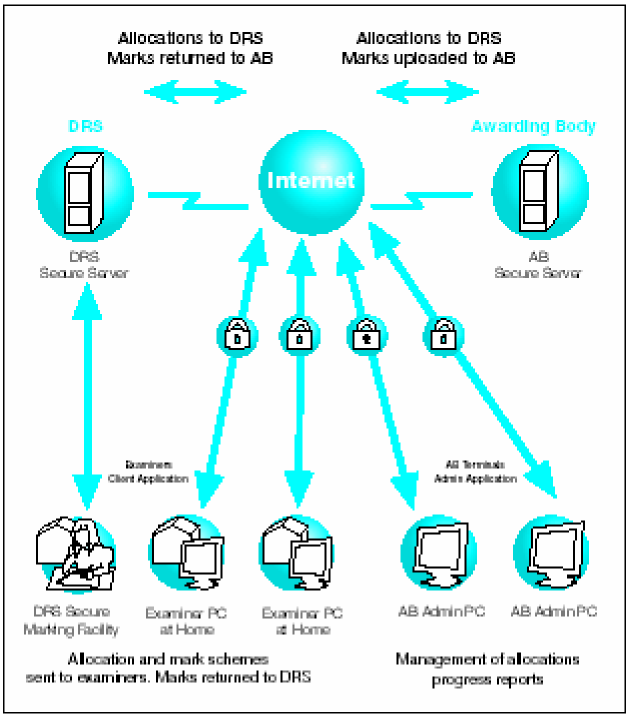
Description of e-Marker® applications

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

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



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.

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.