How effective is fast and automated feedback to examiners in tackling the size of marking errors?

Dr Elizabeth Sykes Independent Consultant in Cognitive Assessment, **Dr Nadežda Novaković, Dr Jackie Greatorex, John Bell, Rita Nádas and Tim Gill** Research Division

Introduction

Reliability is important in national assessment systems. Therefore there is a good deal of research about examiners' marking reliability. However, some questions remain unanswered due to the changing context of e-marking[1], particularly the opportunity for fast and automated feedback to examiners on their marking. Some of these questions are:

- will iterative feedback result in greater marking accuracy than only one feedback session?
- will encouraging examiners to be consistent (rather than more accurate) result in greater marking accuracy?
- will encouraging examiners to be more accurate (rather than more consistent) result in greater marking accuracy?

Thirty three examiners were matched into 4 experimental groups based on severity of their marking. All examiners marked the same 100 candidate responses, in the same short time scale. Group 1 received one session of feedback about their accuracy. Group 2 received three iterative sessions of feedback about the accuracy of their marking. Group 3 received one session of feedback about their consistency. Group 4 received three iterative sessions of feedback about the consistency of their marking. Absolute differences between examiners' marking and a reference mark were analysed using a general linear model. The results of the present analysis pointed towards the answer to all the research questions being "no". **The results presented in this article are not intended to be used to evaluate current marking practices. Rather the article is intended to contribute to answering the research questions, and developing an evidence base for the principles that should be used to design and improve marking practices.**

Background

It is imperative that General Certificates of Secondary Education (GCSE) examinations are marked validly, reliably and accurately. In this article the effectiveness of potential procedures for providing fast and automated feedback to examiners about their marking is evaluated.

For many years a great deal of research resource has focused on the reliability of marking and factors which influence the reliability of marking. The literature covers marking of academic, professional and vocational examinations, as well as marking the work of candidates of varied ages. Examples of research in the field are: Greatorex and Bell (2004; 2008), Akeju (2007), McManus *et al*. (2006), Baird (1998), Richards and Chambers (1996), Williams *et al*. (1991), Laming (1990), Braun (1988), Murphy (1979; 1982). Some, but not a great deal, of this literature focuses on GCSE marking, for example, Suto and Nádas (2007) and Vidal Rodeiro (2007). There are still some unanswered research questions about the effectiveness of different types of examiner training or feedback to examiners in the GCSE context. One such area is the effectiveness of fast and automated feedback to examiners about their marking. With this in mind, the research literature and current practice were used here to develop different approaches to providing feedback to examiners. Subsequently, the effect of each approach on marking accuracy was investigated.

Before setting out the context and the basis of the experimental approaches to feedback, some current pertinent GCSE examining practices need to be noted. For conventional paper-based marking at the beginning of the marking session, examiners normally attend a standardisation meeting. The aim of the meeting is to smooth the progress of high quality marking. In the meeting, scripts and the mark scheme are discussed. After the meeting, examiners submit some of their marked scripts to a senior examiner who reviews their marking and provides individualised feedback to each examiner. Usually the medium of communication is a standard paper form with hand written entries. The form includes marks given by the examiner, the marks given by the senior examiner for the same candidates, and any discrepancies. In some cases the hand written entries provide advice about how to improve marking. Sometimes other supplementary means of communication such as a telephone conversation are used as necessary. If the marking is sufficiently in line with the senior examiner's marking, the senior examiner allows the examiner to continue to mark as they have done so far. If the marking is not sufficiently in line with the senior examiner's, then the process outlined above is repeated. Depending upon the quality of marking, the examiner might not be allowed to mark any further scripts in that examination session. During the marking session further scripts marked by the examiners are sampled and the marking is checked by Team Leaders or the Principal Examiner[2], but feedback is not provided to the examiners. There are also other processes in place for quality control purposes, such as checking marking of scripts near to grade boundaries once grade boundaries have been set; see QCA (2008) for full details.

For each examination there is a range of marks around the Principal Examiner's (PE) or Team Leader's (TL) marking known as 'tolerance'. For many examinations, if an examiner does not mark within tolerance, then they are not an acceptable examiner. However, for some examinations, particularly those including essays, if the examiner's marking is outside

1 E-marking is used here to mean the marking of digital images of examination responses by examiners working at computers.

2 Principal Examiners generally write question papers and are responsible for leading the marking; Team Leaders also oversee some marking.

tolerance but is highly consistent, then the examiner's marking can be scaled. Scaling is the process of adding or subtracting a number of marks from the examiner's marking to bring it in line with the senior examiner's marking. When an examiner is scaled they might be scaled for the whole of the mark range or on part of the mark range. For instance, an examiner might be generous at the top of the mark range and accurate for the rest of the mark range. In which case the marks they gave for most of the mark range would remain unchanged but marks they gave at the top of the mark range would have some marks deducted. During the scaling process the rank ordering of the marks is preserved. One of the few research articles about scaling is Adams and Wilmutt (1982).

For e-marking the process of examiner standardisation is somewhat different to that of conventional paper-based marking. Senior examiners meet to mark a minimum of 35 scripts and agree on what are known as 'definitive marks' for these scripts. The examiners mark a practice sample of scripts remotely. The definitive marks and associated annotations are available for the examiners to consult. Subsequently, the examiners mark ten scripts (standardisation scripts) and submit their marking. Once the marking has been submitted the software informs the examiner of the definitive item level marks for each script. They also receive feedback on their marking from a senior examiner. If an examiner's marking is acceptable they are allowed to go ahead and mark the rest of their allocation. If the marking is not acceptable they can revisit the original standardisation scripts; they also mark another ten scripts and receive feedback on their marking from a senior examiner. If after this second round of feedback the examiner's marking is acceptable, the examiner is cleared by the senior examiner to go ahead and mark the rest of the allocation. If their marking is not acceptable then they are not cleared to continue marking. Once the marking is underway examiners are provided with feedback and monitored. This is accomplished by every 20th script that the examiner marks being a 'seeded script', that is a script for which there is a definitive mark. The differences between definitive marks and examiners' marks can be monitored. If the marking of a seeded script is unacceptable then the Team Leader can review the marking of the last 20 scripts and ask the examiner to re-mark them. The e-marking procedure for standardising marking is different to the conventional paper-based approaches, as feedback can be provided throughout the e-marking session.

There is a wide ranging literature about training and feedback to examiners, much of which is about marking on paper. It is likely that much of the research about paper-based marking is relevant to e-marking. As already noted by Greatorex and Bell (2008), e-marking and linked innovations are associated with the prospect of Awarding Bodies up-dating their practices. In an automated environment, there is the possibility of introducing new training and feedback approaches. For instance, there is the possibility of providing feedback to examiners more quickly than relying on the post. What is more, there is the possibility for the feedback to be automated rather than involving a person-to-person aspect, for example, telephone calls or a face-to-face element, such as co-ordination/standardisation meetings. Bearing these possibilities in mind, our article is intended to investigate which would be the best approach to providing feedback to examiners in an automated environment, based on research evidence.

The traditional reasoning which underpins current paper marking practice is that after examiners have had one or, in some cases, two rounds of feedback and their marking is deemed acceptable, the examiners should continue to mark. It is argued that if they have further feedback then their marking behaviour might change part way through the marking session which makes scaling untenable. There is research that indicates that when conventional paper marking practices are followed some examiners still drift a little over time in terms of their leniency or severity (Pinot de Moira *et al*., 2002). This finding is consistent with other research from beyond the GCSE and A-level context; see Aslett (2006) for a summary. Another argument associated with this traditional line of reasoning is that if feedback is given part way through the marking session the examiners can overcompensate by swinging from severe to lenient or vice versa. This view is also supported by research from outside the GCSE or A-level context such as Shaw (2002), Hoskens and Wilson (2001), as well as Lumley and McNamara (1993). This would then make scaling untenable (unless Awarding Bodies know when responses are marked and are happy to apply different levels of scaling at different times as necessary). In e-marking it is possible to provide feedback iteratively during the marking session. However, this approach contradicts the traditional reasoning.

In some research about feedback to examiners the feedback has been provided shortly after the marking had taken place, perhaps within 24 hours, for example, Hoskens and Wilson (2001). This highlights a limitation of some of the other research in this area such as Shaw (2002) and Greatorex and Bell (2008) where the feedback was received by post and so there was some delay between the marking and receiving feedback.

Another line of traditional reasoning is that examiners should be encouraged either to replicate the marking of the senior examiner, or to be consistently more lenient or severe than the senior examiner. This latter practice is maintained so that examiners can be scaled. The research literature suggests that training or feedback aimed at getting the examiner to be self-consistent (increasing intra-examiner consistency) is likely to be more successful than feedback or training which encourages the examiners to replicate the senior examiner's marking (increasing examiner accuracy or "inter-examiner reliability") (Weigle, 1998; Lunz *et al*., 1991).

To our knowledge some of these issues have not been investigated in the GCSE context. With this in mind the following questions arise:

1) will iterative feedback result in greater marking accuracy than only one feedback session?
2) will encouraging examiners to be consistent (rather than more accurate) result in greater marking accuracy?
3) will encouraging examiners to be more accurate (rather than more consistent) result in greater marking accuracy?

Method

Design

Interventions

This marking experiment applied combinations of four types of interventions:

- examiners receiving one round of feedback
- examiners receiving iterative feedback
- examiners receiving 'accuracy feedback'
- examiners receiving 'consistency feedback'

Each type of intervention is explained in more detail below.

One round of feedback

Examiners received one round of feedback on their marking near the beginning of the marking session.

Iterative feedback

Examiners received feedback on their marking at regular intervals during the marking session.

'Accuracy feedback'

This type of feedback drew examiners' attention to differences between their marks and the reference marks (the reference marks were taken to be the true score for this experiment, more details are given below). The differences between the reference marks and the examiners' marks were provided as *actual differences*. That is, the examiners could see whether the differences were positive or negative and therefore whether they were more lenient or severe than the reference mark. The feedback was presented in graph form so that examiners could see how accurate they were across the entire mark range.

'Consistency feedback'

Examiners received feedback that drew their attention to those responses where the mark they had given deviated in some way from their usual marking level (for example, if they showed a tendency to be in line with the PE or lenient, their attention was drawn to those responses where they marked more harshly). The feedback was presented in graph form so that examiners could see how consistent they were across the entire mark range. In this way, drawing their attention to differences between their marks and the reference marks was avoided, as this could potentially sway the examiners in their marking.

For both the 'accuracy feedback' and 'consistency feedback' interventions, the examiners received written detailed instructions on how to interpret the graphs, before marking began (see Appendix 1). As far as possible the instructions were the same for all groups. The examiners were also given ample opportunity to get in touch with the research team both before and during the marking to raise any queries about the feedback they received. This process was intended to simulate an automated system of providing feedback to examiners on their marking.

During the marking phase, each examiner marked a total of 100 paper responses to one question. The examiners were asked to mark at the item level rather than at the script level because this approach reflects an e-marking environment, where examiners might mark assigned questions rather than assigned scripts (whole question papers).

The four groups marked the same batch of scripts in the same order. There were 5 batches, each consisting of 20 responses covering a wide range of marks. Each batch included the same number of responses in order to avoid a practice effect influencing the accuracy of the marking at each stage in the experiment. Examiners marked one batch of responses per day. The examiners marked the first batch on day one and repeated this exercise with the consecutive batches over each of the following 4 marking days. They received the feedback on their marking (as appropriate) the following morning. Table 1 illustrates the experimental design used in the study.

The first set of 20 responses constituted a practice sample which served as a 'warm-up' exercise to help the examiners remind themselves of the mark scheme and prepare them for marking the remaining four sets of responses. Thus, no group received any feedback after marking the first batch.

Table 1: Experimental design

Day	*Accuracy feedback*		*Consistency feedback*	
	Group 1	*Group 2*	*Group 3*	*Group 4*
1	Batch 1	Batch 1	Batch 1	Batch 1
2	Batch 2	Batch 2	Batch 2	Batch 2
3	*Feedback on batch 2*	*Feedback on batch 2*	*Feedback on batch 2*	*Feedback on batch 2*
3	Batch 3	Batch 3	Batch 3	Batch 3
4		*Feedback on batch 3*		*Feedback on batch 3*
4	Batch 4	Batch 4	Batch 4	Batch 4
5		*Feedback on batch 4*		*Feedback on batch 4*
5	Batch 5	Batch 5	Batch 5	Batch 5

Procedure

Only one question was selected to be used in the research. After live marking responses to that one question in some OCR scripts were copied. All the copies were cleaned of marks. Thus multiple copies of the same responses could be marked by many examiners.

Two PEs were asked to give their own reference marks for candidates' responses. The PEs then compared their marks and agreed on a reference mark for each response. This approach reflects the procedures used to determine definitive marks in an e-marking context/ environment.

Each experimental group (1 to 4) experienced the interventions as described above. All the marking was undertaken remotely. Examiners were expected to spend around 120 minutes marking each batch (it takes approximately five minutes or less to mark the question). The 5 batches of responses were sent out to examiners by post. After marking a batch the examiners sent their marks back to the research team by e-mail, and received the feedback by e-mail.

Script samples

A GCSE English Higher Tier examination question was used in the experiment. Candidates could score 30 or fewer marks on the question. A total of 100 responses with reference marks were divided into 5 batches of 20 responses. Each batch of 20 responses was intended to include a similar range of achievement. The resulting frequency of reference marks by batch is given in the Table 2 below.

Participants

In addition to the two PEs a total of 33 examiners took part in the study. All the examiners were experienced examiners who had marked the GCSE English Higher Tier examination in live marking. Other reasons for recruiting these particular examiners included that they were all contactable by email and available to mark at the scheduled times. The examiners were divided into four experimental groups: two groups consisted of nine examiners, one group consisted of eight examiners and one group of seven examiners. The differences in numbers in groups were due to issues like availability and dropout.

To form the groups, the examiners were matched in terms of their

Table 2: Reference marks (agreed by 2 PEs) and the frequency of each reference mark in each batch

Reference mark	*batch 1*	*batch 2*	*batch 3*	*batch 4*	*batch 5*
13	0	0	0	0	1
14	0	0	1	1	0
15	2	2	1	1	1
16	0	0	1	1	2
17	2	2	2	1	0
18	2	3	1	2	2
19	3	2	3	3	3
20	3	1	2	0	2
21	0	2	1	3	1
22	1	2	3	2	3
23	3	1	1	1	2
24	1	2	1	2	0
25	1	2	2	1	0
26	0	0	0	0	1
27	1	0	0	1	1
28	0	1	1	0	1
29	0	0	0	1	0
30	1	0	0	0	0

Table 3: Number of participants who were lenient or severe in a previous session of live marking of the examination

Group	*Neither lenient or severe*	*Lenient*	*Severe*	*Total*
1	3	5	1	9
2	4	5	0	9
3	3	4	1	8
4	2	4	1	7

severity in the previous live marking session; the intention was that there would be a similar variety of marking severity in each group to avoid group differences.

Table 3 provides a summary of the final distribution of the historical severity and leniency of examiners who went on to complete all aspects of the study. For the purposes of Table 3 examiners were classified according to their live marking of the examination in the previous live marking session. The classifications were 'neither lenient nor severe' if they were not scaled, 'lenient' if their scaling resulted in marks being deducted and 'severe' if their scaling resulted in marks being added.

Results

A statistical analysis of the absolute differences between the examiner's mark and the reference mark was conducted. When we discuss the analysis and results from our data in this article we refer to absolute differences as a measure of accuracy[3] or marking error. To report *absolute differences* all negative differences were changed to positive values. This has the advantage that the overall size of the marking error can be seen, regardless of the levels of the severity or leniency of marking. Reporting absolute differences also has the advantage that a lower mean absolute difference is an improvement in accuracy, whereas this information is lost when reporting actual differences (where positive and negative differences can negate each other).

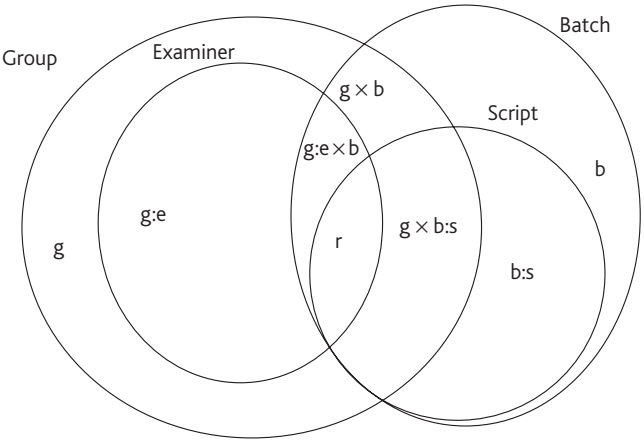

Figure 1: Diagram representing the structure of the experiment

Figure 1 above represents the structure of the experiment. In the diagram 'g' represents experimental group, 'e' represents examiner, 'b' represents batch, 's' represents response and 'r' represents residual error. Examiners are nested within groups (g:e) crossed with responses nested within batches (b:s). This indicates that it is possible to estimate two main effects, batch and group, and five interaction effects (examiner within group, response within batch, group crossed with batch, examiner within group crossed with batch and a group crossed with response within batch). Finally, there is a confounded residual error. Ideally there should be no differences between groups. Examiners within groups and responses within batches are expected to be different. Batch and group crossed with batch are effects that the experiment was designed to estimate.

The above model can be represented as an equation:

$$y_{gebs} = \mu + \tilde{\mu_g} + \tilde{\mu_b} + \tilde{\mu_{g\times b}} + \tilde{\mu_{g:e}} + \tilde{\mu_{b:s}} + \tilde{\mu_{g:e\times b}} + \tilde{\mu_{g\times b:s}} + \tilde{r_{gebs}}$$

where y_{gebs} is the marks difference for examiner e in group g marking response s in batch b,

- μ is the grand mean,
- $\tilde{\mu_g}$ is the overall effect of group g,
- $\tilde{\mu_b}$ is the overall effect of batch b,
- $\tilde{\mu_{g\times b}}$ is the effect of the interaction between batch b and group g,
- $\tilde{\mu_{g:e}}$ is the effect of examiner e in group g,
- $\tilde{\mu_{b:s}}$ is the effect of response s in batch b,
- $\tilde{\mu_{g:e\times b}}$ is the effect of the interaction between examiner e within group g crossed with batch b,
- $\tilde{\mu_{g\times b:s}}$ is the effect of the interaction between group g crossed with response s within batch b,
- $\tilde{r_{gebs}}$ is the error term.

The foci of this study are the batch effect, the interaction between group and batch and the interaction of examiner within group crossed with batch.

3 The reader might notice that when we discuss accuracy in the context of accuracy feedback we are concerned with actual differences, when we are discussing the analysis of our data we are discussing absolute differences and when we are discussing the research literature we might be referring to actual or absolute differences. These uses of the term accuracy are in keeping with much of the research literature.

Table 4: The General Linear Model

Source	*df*	*Type III SS*	*Mean Square*	*F Value*	*Pr > F*
Group	3	81.84	27.28	5.96	0.0005
Batch	4	166.96	41.74	9.11	<.0001
Examiner (group)	29	2553.45	88.05	19.22	<.0001
Response (batch)	95	7282.17	76.65	16.73	<.0001
Group*batch	12	104.03	8.67	1.89	0.0308
Examiner (group) * batch	114	1993.45	17.49	3.82	<.0001
Group*response (batch)	285	1105.77	3.88	0.85	0.9656
Error	2717	12446.63	4.58		

A general linear model was applied to the absolute differences between the examiner's marks and the reference mark.

The results in Table 4 indicate that most of the effects were significant ([Pr>F]<0.05). The results can be taken to mean that:

- in general the marking of each group was different;
- in general each examiner's marking changed over batches;
- individual examiners within a particular group had different levels of marking accuracy;
- the accuracy of marking was different for different responses;
- each group's marking accuracy changed from batch to batch (generally accuracy was improved over time until batch 5 when marking became more inaccurate);
- the examiners in different groups marked the different batches differently;
- the experimental groups of examiners did not generally mark the same response differently, i.e. the experimental groups tended to have similar accuracy levels for the same response.

Figure 2 illustrates that the marking accuracy of all groups generally increased with each batch except for the final batch of marking. In this analysis least square (LS) means can be used in the way that an arithmetic mean would be used in other situations.

Multiple comparisons procedures, like the general linear model, are used to control for the familywise error rate. For example, suppose that we have four groups and we want to carry out all pairwise comparisons of the group means. There are six such comparisons: 1 with 2, 1 with 3, 1 with 4, 2 with 3, 2 with 4 and 3 with 4. Such a set of comparisons is called a family. If we use, for example, a t-test to compare each pair at a certain significance level $\propto$, then the probability of Type I error (incorrect rejection of the null hypothesis of equality of means) can be guaranteed not to exceed $\propto$ only individually, for each pairwise comparison separately, and not for the whole family. To ensure that the probability of incorrectly rejecting the null hypothesis for any of the pairwise comparisons in the family does not exceed $\propto$, multiple comparisons methods that control the familywise error rate (FWE) need to be used (Westfall *et al*., 1999).

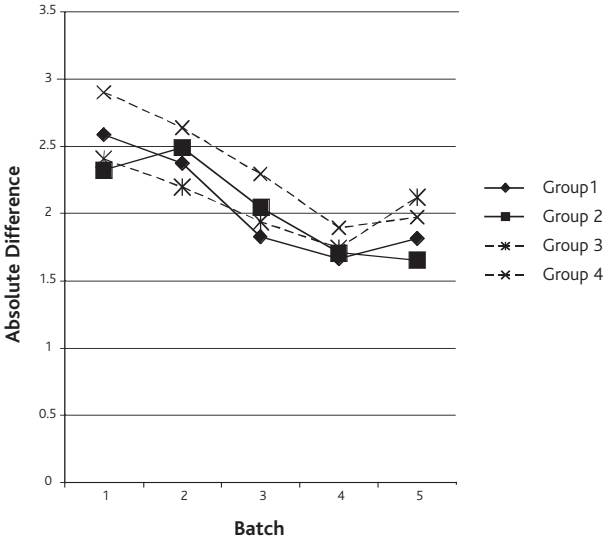

Figure 2: LS means of group by batch

The LS means for the effect of batch are shown in Table 5 and illustrated in Figure 3. Table 6 shows whether the means of each pair of batches are statistically significantly different.

Table 5: Adjustment for multiple comparisons: absolute differences

Batch	*LS mean*	*95% Confidence Limits*	
1	2.55	2.47	2.68
2	2.43	2.30	2.56
3	2.02	1.89	2.15
4	1.75	1.62	1.88
5	1.88	1.74	2.01

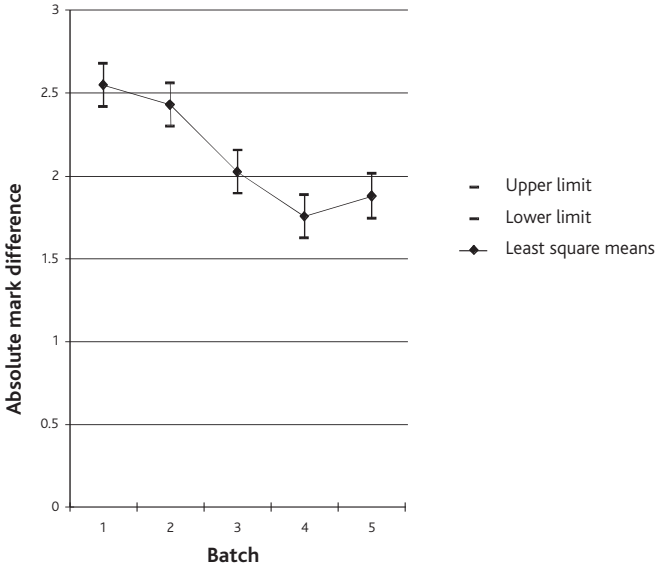

Figure 3: LS means by batch (with confidence intervals) for all groups

Table 6: LS means for the effect of batch: absolute differences

Batches	*t value*	*Pr > (t)*
1–2	1.27	0.71
1–3	5.57	< 0.01
1–4	8.43	< 0.01
1–5	6.99	< 0.01
2–3	4.23	< 0.01
2–4	7.16	< 0.001
2–5	5.74	< 0.001
3–4	2.87	0.03
3–5	1.53	0.55
4–5	1.28	0.71

Overall, the changes in accuracy as measured by mean absolute differences were as follows:

(1) **Batch 1–Batch 2:** there was no feedback provided but there was increased familiarity with the Mark Scheme. There was a slight non-significant improvement in accuracy.

(2) **Batch 2–Batch 3:** all groups received feedback. This comparison showed that there was a significant improvement in accuracy and this was the largest improvement between consecutive batches.

(3) **Batch 3–Batch 4:** Groups 1 and 3 had no feedback, Groups 2 and 4 had feedback. This comparison showed that there was a significant improvement in accuracy and that all groups continued to improve in accuracy.

(4) **Batch 4–Batch 5:** Groups 1 and 3 had no feedback, Groups 2 and 4 had feedback. There was no improvement in accuracy for any group. In fact there was a slight non-significant decline.

Accuracy improved for all groups of examiners after they had the first round of feedback. The improvement was sustained for another round of marking for all groups whether they received continued feedback or not. Performance then levelled off on the last round of marking for all groups.

Thus, in terms of LS means, the findings showed that the first round of feedback (accuracy and consistency) was effective in bringing the examiners' marking nearer to the reference mark and that the difference in the mean marks between examiners and the reference mark was reduced from 2.55 marks to 2.02 marks. There was continued improvement for one more round of marking, reducing the difference in the mean marks from 2.02 to 1.75. The mean mark for every group was within 2 marks of the reference mark by the fourth batch. Improvement appeared to level off at this point although the mean mark difference between examiners and the reference mark for the fifth batch remained below two marks. The pattern was the same for all of the groups, suggesting that initial feedback per se was effective in reducing marking error, but that neither the type nor the amount of feedback were important in contributing to improved accuracy.

It is worth noting that in this analysis the main comparison is between the marking trajectories of the different groups rather than a direct comparison between each group's marking at each stage of the experiment.

Discussion

Awarding Bodies have indicated a keen interest in examiner training in the GCSE context. Advances in computerised technology have provided the opportunity to consider their impact on the possibility of updating methods for providing training and feedback to examiners during the marking sessions. Being able to mark responses on screen and receive feedback by email shortly after each marking session rather than by post might both be expected to impact on the reliability of examiner marking.

The aim of this study was to investigate how feedback might affect levels of reliability of examiners' marking in the GCSE context and to consider the results in the context of an automated environment. The administration of two different amounts of feedback (once and three times) and of two different types of feedback (accuracy and consistency) were investigated.

The accuracy of examiners' marking was investigated by measuring the absolute differences between the examiners' marks and the reference mark. There were significant differences between the four groups of examiners and the five batches of responses. However, all of the groups performed similarly across batches. The marking of all groups improved in accuracy over the course of the study, with the greatest improvement being evident after the first round of feedback. The improvement was sustained for another round of marking for all groups whether they received continued feedback or not. Performance then levelled off on the last round of marking. The mean mark for each group was approximately 2 marks off the reference mark by the fourth batch and remained at this level to the end of the study. The mean mark for all groups together was within 2 marks of the reference mark by the fourth batch and remained so to the end of the study. Thus initial feedback per se was effective in reducing marking error, but neither the type nor the amount of feedback was important in contributing to improved accuracy. In other words our analysis of absolute differences indicated that the answer to all three research questions is 'no'.

Similarly, Shaw (2002) noted increases in accuracy up to batches 3 and 4, although these were not maintained in the fifth batch of marking. By the end of his study, accuracy levels had returned to the level they were at the start of the study. The tailing off in increases in accuracy may have been the result of 'participation fatigue' (Shaw, p. 17). Shaw suggested that the increases in accuracy were the result of feedback but there was no control group to test this theory. Likewise Greatorex and Bell (2008) suggested that feedback could have led to an increase in marking accuracy, but these researchers recognised that, as in Shaw's study, the research design did not include a no-feedback control condition in order to clarify this suggestion. Furthermore, Greatorex and Bell found no clear pattern to suggest which kind of feedback might account for the rise in accuracy. The current study had the benefit of a control group to make identification of an effect (or non-effect) of iterative feedback more discernible.

In Shaw (2002), and Greatorex and Bell (2008), the feedback was not given immediately after the marking had taken place, but it was provided a little later due to providing the feedback by post. Although this reflects some current practice, examiners might benefit from more immediate feedback. In both studies feedback on the previously marked batch was provided just before the next batch was marked. One of the aims of the current study was to provide feedback within 24 hours of marking, as in Hoskens and Wilson (2001).

A limitation of the present study is that not all possible forms of feedback were researched. Arguably, a further limitation of the research concerns the allocation of participants to groups, which was based on the severity of previous live marking at the examination level. The marking in this study is at the item level. It is possible that the severity of live marking at the examination level is not linearly related to severity of experimental marking at the item level, and it is beyond the scope of this article to investigate this relationship. However, the size of the mean marking error for different groups in batches 1 and 2 differs by less than a mark (see Figure 2). This suggests that the groups were fairly well matched at the beginning of the study in terms of the size of the marking error.

There is a caveat for using the results presented in this article, as follows. We analysed only absolute differences and not actual differences between the examiner's mark and the reference mark. For *absolute differences* all negative differences were changed to positive values. This has the advantage that the overall size of the marking error can be seen, regardless of the levels of severity or leniency. Analysis of *actual*

differences between the examiner's mark and the reference mark (negative differences remain negative) provides information regarding levels of severity or leniency. Sometimes the analysis of actual and absolute differences can lead to different research outcomes, one such case in a marking study is one of the experiments reported in Baird *et al*. (2004). However, for this article we were concerned with the accuracy of the marking or the size of marking errors, which is estimated using absolute differences.

It should also be noted that the results presented in this article cannot be used alone to evaluate the utility of current live marking practices. To use the results presented here it is advisable to:

- **investigate how different types of feedback affect severity and leniency which are not considered in this article;**
- **note that the experiment intended to simulate potential procedures for an automated environment and answer research questions, and not to evaluate the utility of live marking practices, which are different to the procedures in the experiment.**

One line of traditional reasoning that underpins current practice is that after examiners have had one (or in some cases two) round(s) of feedback and their marking is acceptable, the examiners should be left to mark. Some research indicates that some examiners drift a little over time in terms of their leniency or severity even with the initial feedback (Pinot de Moira *et al*., 2002). Other research shows that iterative feedback can lead to examiners swinging from leniency to severity or vice versa (Shaw, 2002; Hoskens and Wilson, 2001; Lumley and McNamara, 1993). It was beyond the scope of this article to investigate whether examiners' marking swung from severe to lenient. The analysis of absolute differences in the present article indicated that marking accuracy tended to increase throughout the study (except for the final batch) but that the iterative feedback was no better than one-off feedback in tackling marking errors. Indeed the initial feedback was the most effective; this might be partly because at the beginning of the study there was a greater marking error to rectify. This suggests that there would be no apparent benefit in providing feedback (of the types used in this study) throughout an e-marking session based on absolute differences between examiners' marking and the reference marks.

The other lines of traditional reasoning are that examiners should be encouraged either to replicate the marking of the senior examiner, or to be consistently more lenient or severe than the senior examiner. Previous research suggests that training or feedback aimed at getting the examiner to be consistently severe or lenient in comparison to the senior marker is likely to be more successful than feedback or training to encourage the examiners to replicate the senior examiner's marking (Weigle, 1998; Lunz *et al*., 1991). The analysis of absolute differences did not indicate that one approach was more beneficial than the other.

References

Adams, R.M. & Wilmut, J. (1982). A measure of the weights of examinations components, and scaling to adjust them. *The Statistician*, **30**, 263–9.

Akeju, S. A. (2007). The reliability of General Certificate of Education Examination English composition papers in West Africa. *Journal of Educational Measurement*, **9**, 3, 175–180.

Aslett, H. J. (2006). Reducing variability, increasing reliability: exploring the psychology of intra- and inter-rater reliability. *Investigations in University Teaching and Learning*, **4**, 1, 86–91.

Baird, J. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research*, **40**, 2, 191– 202.

Baird, J., Greatorex, J. & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education Principles, Policy and Practice*, **11**, 3, 331–348.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational and Behavioural Statistics*, **13**, 1, 1–18.

Greatorex, J. & Bell, J. F. (2004). Does the gender of examiners influence their marking? *Research in Education*, **71**, 25–36.

Greatorex, J. & Bell, J.F. (2008). What makes AS Marking Reliable? An Experiment with some stages from the Standardisation Process. *Research Papers in Education*, **23**, 3, 333–355.

Hoskens, M., & Wilson, M. (2001). Real-Time Feedback on Rater Drift in Constructed-response Items: An Example from the Golden State Examination. *Journal of Educational Measurement*, **38**, 2, 121–145.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgements. *The Quarterly Journal of Experimental Psychology Section A*, **42**, 2, 239–254.

Lumley, T. & McNamara, T.F. (1993). Rater Characteristics and Rater Bias: Implications for Training. *Language Testing*, **12**, 54–71.

Lunz, M.E., Stahl, J.A. & Wright, B.D. (1991). *The invariance of judge severity calibrations*. Paper presented at the annual meeting of the American Research Association, Chicago, IL. Quoted in Weigle, S.C. (1998). Using FACETS to Model Rater Training Effects. *Language Testing*, **15**, 2, 263–287.

McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education*, **6**, 42 http://www.biomedcentral.com/1472–6920/6/42

Murphy, R. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, **52**, 1, 58–63.

Murphy, R. (1979). Removing the marks from examination scripts before re-marking them: does it make a difference? *British Journal of Educational Psychology*, **49**, 1, 73–78.

Pinot de Moira, A., Massey, C., Baird, J. & Morrissy, M. (2002). Marking consistency over time. *Research in Education*, **67**, 79–87.

Qualifications and Curriculum Authority (2008). *GCSE, GCE and AEA code of practice 2008*, (London, Qualifications and Curriculum Authority) http://ofqual.gov.uk/files/Code_of_practice_April_2008.pdf

Richards, B. & Chambers, F. (1996). Reliability and validity in the GCSE oral examination. *Language Learning Journal*, **14**, 28–34.

Shaw, S (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes*, **8**, 13–17.

Suto, I. & Nádas, R. (2007). The 'Marking Expertise' projects: Empirical investigations of some popular assumptions. *Research Matters: a Cambridge Assessment Publication*, **4**, 2–5.

Vidal Rodeiro, C. L. (2007). Agreement between outcomes from different double marking models. *Research Matters: a Cambridge Assessment Publication*, **4**, 28–34.

Weigle, S.C. (1998). Using FACETS to Model Rater Training Effects. *Language Testing*, **15**, 2, 263–287.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D. & Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*, SAS Institute.

Williams, R., Sanford, J., Stratford, P. W. & Newman, A. (1991). Grading written essays: a reliability study. *Physical Therapy*, **71**, 9, 679–686.

APPENDIX 1:
INSTRUCTIONS FOR INTERPRETING CONSISTENCY FEEDBACK FOR GROUP 3

Dear examiner,

This document is intended to prepare you for the feedback you will receive after marking Batch 2. It contains explanations as to what the feedback will look like and how to interpret it. Please read this before you start marking. If you have any questions or are unclear about anything related to the feedback you will receive, please do not hesitate to contact us as soon as possible.

The feedback you will receive will be different from the feedback you receive in live marking (after standardisation sample). In live marking, the feedback you receive shows the difference between your marks and Principal Examiner's marks. However, the feedback you will receive here will show the extent to which the marks you have given to certain responses differ from your average marking for that specific mark range. In other words, the feedback will not focus on how different your marking is from that of the PE, but it will focus on the consistency of your marking.

You will receive feedback on all the marks you have given to responses within a batch. The feedback you receive will be in the form of a graph similar to the graph presented below (these are made-up data).

As you can see, the graph consists of two axes. The X-axis is a thick horizontal line running through the middle of the graph. The ticks on this line represent marks, from 0 to 30, that can be given to a candidate's work.

The Y-axis is the leftmost vertical line and it shows to what extent your marks differ from your average marking. If this difference is above 0, this means that you have marked a candidate's work more generously than would be expected if your average marking is taken into account. If the difference is negative, i.e. below 0, this means that you were harsher than would be expected if your average marking is taken into account.

The "diamonds" scattered over the graph plot area represent candidates' responses from the batch. These are marked as r1 (response 1), r2 (response 2) etc. and refer to the number on the first page of each candidate's response, which is also the number in the mark recording sheet which we will send you to record your marks.

Let us take, for example, responses number r6 and r14. If you traced an imaginary line from the "diamond" representing script r6 onto the horizontal X-axis, it would cross it at 20, showing that you have given this response a mark of 20. If you traced an imaginary line onto the vertical Y-axis, it would cross it at close to +3, indicating that the mark you gave to this candidates' work was about three marks higher than your average marking. In other words, if your average marking is taken into account, we would have expected you to have given this response r6 a mark of 17, rather than 20. On the other hand, the mark you gave to candidate response number 14 (r14) is consistent with your average marking for this mark range.

The more clustered your marks are around the X-axis, the more consistent you are in your marking for that specific mark range. The more spread out your marks are, the more inconsistent you are in your marking. Furthermore, by taking a look at the graph as a whole you can get an overall impression as to the overall spread of your marks.

We will email you feedback as part of an attached Microsoft Excel sheet.