Centre for Education Research and Policy

# Why Item Mark?

## The Advantages and Disadvantages of E-Marking

Anne Pinot de Moira

## Introduction

The move towards e-marking has meant some large changes to the long established processes and practices for marking national examinations. Each change has been justified with a business plan and, while there are clear benefits to the awarding bodies, the benefits to examiners and candidates have not always been quite so obvious. By considering the advantages and disadvantages, this report explains why a move to e-marking ultimately provides more reliable marking (see Further Reading: [1]).

## The Administrative Effect

Without doubt there are some administrative drawbacks to the move towards e-marking, but most of these are only short-term drawbacks due mostly to the novel nature of the task. In order to e-mark, examiners need to have access to a computer and to a reliable broadband internet connection. They have to learn how to use a new interface while maintaining the same high level of marking reliability. They are exposed to new feedback mechanisms and new rules surrounding the marking and they must become accustomed to working exclusively at a computer screen.

On the other hand, a number of administrative burdens have been removed. There is no longer the need to accommodate vast numbers of paper scripts or be available to receive these scripts from the postal service. During the marking period, there is no need to post scripts for sample re-marking or to return them to the awarding body at the end of the marking period. This decreases the time in transit, limits the risk of scripts being lost and, at a more superficial level, allows for a less cluttered marking environment. Furthermore, the collection of marks using an e-marking system eliminates most of the clerical errors associated with paper-marked scripts.

- The introduction of e-marking has some short-term drawbacks related to the learning of new processes and practices.
- The electronic working environment eases time constraints in the examining period and eliminates most clerical errors.

## The Effect on Marking Reliability

### Examiner Idiosyncrasies

All examiners are different and they naturally look upon the work of individual candidates differently. Standardisation meetings are held to create a common view of the quality of work presented. Unfortunately, there will usually be some degree of subjectivity that remains in the decision making process. In the past, sample re-marking has been used to check and adjust for any examiner idiosyncrasies. Following the introduction of e-marking, although sample re-marking still exists as a quality assurance measure to ensure reliability, the need for examiner adjustment has been removed. This is due to the fact that splitting the marking of a single script

The Assessment and Qualifications Alliance (AQA) is a company limited by guarantee registered in England and Wales
(company number 3644723). Registered address: AQA, Devas Street, Manchester M15 6EX

across more than one examiner has the inherent advantage of reducing the effect of individual idiosyncrasies and the need to make adjustments.

*A Little Bit of Statistics*

For the sake of illustration, let us assume that each item in a test has its own true mark. Let us also assume that any deviation from this true mark is called error. If examiner standardisation has been effective and every examiner has marked a given item, then the distribution of errors around the true mark for this item should be symmetrical. In other words, there should be as many examiners regarding the item as better than the true mark as there are examiners who consider it to be worse than the true mark. From this assumption, we can devise a probability distribution of the deviation from the true mark for that item, as shown in Figure 1.

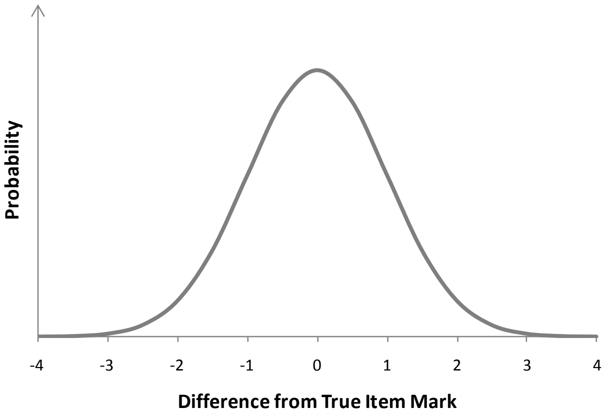

**Figure 1 Probability distribution of deviation from the true item mark**

It is possible to add the error distribution for each item on a script together to arrive at an error distribution for that script as a whole. If one examiner is marking the whole script, it is safe to assume that there will be some relationship between the degree of leniency (or severity) shown on one item with that shown on another item (see Further Reading: [2]). On the other hand, if several examiners contribute to the final mark of a complete paper, as happens with e-marking, then there will probably be little or no relationship between items in terms of deviation from the true mark.

In a ten item examination, Figure 2 shows the cumulative effect on reliability of the errors on each item. The horizontal axis represents the difference from the true mark, expressed as a percentage of the maximum mark for the script. The vertical axis represents the probability of the true mark being awarded. If you look at the peak in the middle of the graph, you can see that the probability of being awarded the true mark for the script is far higher when ten examiners are involved in the marking, than when there is only one examiner. Also, there are fewer candidates to be found at the ends of the distribution curve with ten examiners. In fact, simply having more than one examiner contributing to the marking of a single script is better, even when there are only two items to be marked, as shown in Figure 3.

It could of course be argued, on the basis of Figure 2 and Figure 3, that the effect of having multiple examiners marking a single script would be to reduce the standard deviation of the distribution of marks for the examination. Thus discrimination between candidates might be compromised. However, any reduction in the spread of marks as a result of item marking is a consequence of reduced error. So any decrease in discrimination would, in fact, be a decrease in spurious discrimination.

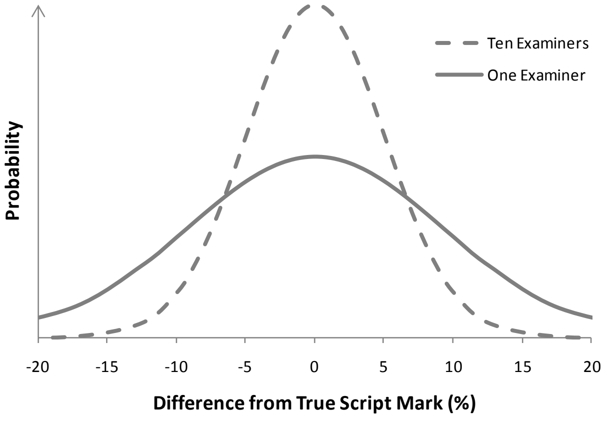

**Figure 2 Probability distribution of deviation from the true script mark (ten item script)**

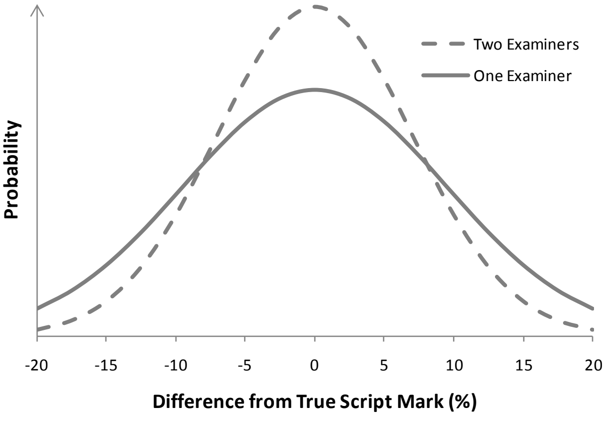

**Figure 3 Probability distribution of deviation from the true script mark (two item script)**

In real life, the true mark is a debatable concept and so an error distribution is not something that can be observed readily and, even if it could be, it would not be a smooth, elegant curve like the one in Figure 1. Although the theory which can be applied to reveal the patterns shown in Figure 2 and Figure 3 is fairly robust to using real data, it is still important to realise that examiner standardisation must be structured to minimise any systematic, or non-random, bias in deviations from the true mark, however the true mark is defined.

- Item marking will always provide more reliable script marks than whole paper marking.
- It is important that examiner standardisation minimises systematic marking biases.

**Examiner Biases**

*The Halo Effect*

Examiner biases do not always have a negative influence on the reliability of marking. For example, it may sometimes be quite valid to refer to a candidate's earlier response to help judge the current work, such as when awarding follow-through marks to give credit to logic, rather than just to the final answer.

At other times, prior information about a candidate might not be useful. For example, an examiner might carry forward preconceived, and possibly incorrect, ideas about a candidate's understanding based on answers to previous, yet unrelated, items. In an examining context, this bias based on prior information is often called a halo effect (see Further Reading: [3], [4]). A halo effect might be observed within the marking of a single script but it might also be identified across scripts, between candidates.

An effective system of item marking allows the positive influence of the halo effect to remain while removing the negative aspects. Related items are presented as connected (clipped) so that, where previous answers should have a bearing on the mark awarded, the halo effect is encouraged. On the other hand, where previous answers are unrelated, the examiner can mark the item without irrelevant or unhelpful preconceptions.

Even with e-marking there is still the potential for a halo effect between items presented sequentially but, as items are presented in a random order, the bias is not systematic. Any bias affecting one item on a script will be different from that affecting another item on the same script. The facility to revisit items marked previously also allows examiners to check that they are marking consistently.

*Other Biases*

In an e-marking system, each item is completely anonymous, so there is no risk of the script origin influencing expectations, simply because an examiner has no idea of the candidate's name, centre of entry, gender or ethnicity (see Further Reading: [5]). Having said that, handwriting remains a feature which could introduce bias to the marking (see Further Reading: [6]). Nevertheless, the fatigue associated with marking an item with poor handwriting might be lower compared with marking a whole script, reducing any inadvertent bias on that item.

- A system of e-marking allows valid halo effects while reducing the influence of, possibly incorrect, preconceived ideas about a candidate's ability.
- E-marking is as good, or better, than paper-based script marking at reducing unwanted biases.

### Examiner Expertise

Deconstructing a script into its constituent items allows marking to be distributed in line with examiners' marking skills and abilities. For example, there is an obvious advantage to being able to direct low tariff items with objective answers towards examiners with more general marking skills. Higher tariff items that have subjective responses can then become the sole focus of examiners who have subject expertise. Furthermore, assuming that examiner standardisation is adequate, in specifications which have optional items requiring knowledge in specific areas of the subject, examiners who have this specific knowledge can then be used.

- Item level marking allows the best use of subject expertise.

## The Effect of Real-Time Quality Assurance

In a paper-based scheme, where sample re-marking is in place, the examiner provides evidence at a single point in time. The examiner chooses the evidence to send and, beyond the first training sample, no feedback on the quality of marking is given based on this evidence. The sample re-marking system is used to determine whether an examiner is fit to continue marking and, if so, whether his or her marking requires adjustment.

Unlike paper-based marking, e-marking makes it possible to monitor the performance of examiners throughout the marking period. Examiners have to show on a daily basis that they are 'fit' to mark an item by qualifying. In order to qualify, they must successfully mark a set of responses which have been pre-marked by a senior examiner. The definition of success will depend upon parameters set for the item.

For short answer items, after qualification, marking problems are identified by the introduction of pre-marked seeds into an examiner's allocation. Failing to mark correctly a predetermined number of seeds in a set time period means that an examiner is temporarily stopped from marking that item. The procedure is similar for longer essay-style items, except that marking problems are identified by observing discrepancies in double-marked items.

Every time an examiner is stopped from marking, that examiner must discuss issues surrounding the marking of that item with a member of the senior examining team before marking can recommence. So the advantage of real-time quality assurance over sample re-marking is the facility to provide immediate performance feedback to examiners. The parameters set to stop marking are adjustable and can be adapted to suit the characteristics of an item, resulting in continual improvements to quality assurance. Furthermore, item marking removes the need to adjust examiners' marks.

Any quality assurance process which monitors human activity is open to manipulation. For example, there is much anecdotal evidence that the sample of scripts sent for re-marking in the paper-based system was never entirely randomly selected by examiners. So too, in an e-marking system, there might be ways of playing the system but the level of flexibility in the control mechanisms means that such loopholes can be closed more easily. Nevertheless, any system in which subjective judgement is used relies, to a certain extent, on trust. So it is important for examiners to understand that, although disconcerting at times, quality assurance measures are in place to ensure continual improvement, and the measures are ultimately for the greater good.

- Compared with paper-based sample re-marking, real-time quality assurance enables a focus on continual improvement rather than the post-hoc adjustment.
- Examiners need a clear understanding of the motivation behind quality assurance measures.

## Further Reading

[1] D. Fowles, “Literature review on effects on assessment of e-marking,” Manchester: AQA Centre for Education Research and Policy, 2005.

[2] A.M. Pinot de Moira, “Marking consistency over time,” Manchester: AQA Centre for Education Research and Policy, 2002.

[3] M. Spear, “The influence of halo effects upon teachers' assessments of written work,” *Research in education.*, 1996, p. 85.

[4] M. Spear, “The influence of contrast effects upon teachers' marks,” *Educational research.*, vol. 39, 1997, p. 229.

[5] J. Baird, “Bias in Marking,” *AQA Internal Report*, 1995.

[6] L.R. Markham, “Influences of Handwriting Quality on Teacher Evaluation of Written Work,” *American Educational Research Journal*, vol. 13, 1976, pp. 277-283.