Quality of Marking

Review of Literature on Item-level Marking Research

February 2014

Ofqual/14/5378

Contents

Introduction

This report presents the findings of a literature review on the possible benefits of item-level marking over whole-script marking.

In our interim report *Review of Quality of Marking in Exams in A levels, GCSEs and Other Academic Qualifications*,[1] we reported that, in summer 2012, about two-thirds of all scripts for GCSE, A level and other academic exams were scanned into digital format and sent to examiners for marking on a computer screen, via a secure system. This type of marking enables examiners to mark at item level, where a scanned script is split up into individual questions (or groups of related questions), which are marked by different examiners.

Since its introduction by Pearson Edexcel in 2003, on-screen marking has grown rapidly and is now used by all exam boards to some degree. However, not all exam boards use item-level marking. AQA, Pearson Edexcel and WJEC CBAC use item-level marking for many of their scripts that are marked on-screen, whereas the Council for the Curriculum, Examinations and Assessment (CCEA), OCR, the International Baccalaureate (IB) and Cambridge International Examinations (CIE) only use whole-script marking, where each exam script is marked by a single examiner. In summer 2012, just over half of all exam scripts[2] across these seven exam boards were marked as whole scripts (54 per cent), rather than split into items (45 per cent).

Background

Before we started this literature review, we commissioned the National Foundation for Educational Research to conduct a literature review on marking reliability research, which included literature on item-level marking. In March 2013, we subsequently interviewed senior representatives from the seven exam boards listed above on marking quality, including their approach to item-level marking and the rationale behind their marking methods. The outcomes of these pieces of work relevant to this literature review are summarised here.

Findings from the literature review on marking reliability research

The National Foundation for Educational Research report *A Review of Literature on Marking Reliability Research* (Tisi *et al*., 2013) identified a number of ways in which

[1] www.ofqual.gov.uk/files/2013-06-07-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-interim-report.pdf

[2] This includes scripts marked online and traditionally.

item-level marking could improve marking reliability and the quality of marking, including:

- Reducing the effect of bias caused by the rest of the exam paper. When one examiner marks all the questions on an exam script, the mark he or she allocates to one item may be affected by the student's responses to other unrelated questions. In other words, an examiner might carry forward preconceived ideas about the level of a student's understanding (whether positive or negative assumptions) based on answers to previous unrelated items. This is known as the halo effect (Spear, 1996, cited in Meadows and Billington, 2005). When different examiners mark each item on a script, random errors in their marking are likely to cancel each other out. So, for every question that is over-marked there is likely to be one that is under-marked. This means the more examiners who contribute to the final mark of a script; the more reliable the final mark will be, reducing the influence of a single examiner on an exam script (Pinot de Moira, 2011).

- Enabling distribution of questions to examiners with the appropriate level of expertise. For example, items with a range of acceptable answers that can be fully defined could be marked by individuals who do not necessarily have the experience to mark more complex items. Items that require more complex marking strategies could be marked by the most experienced examiners (Pinot de Moira, 2011; Meadows and Billington, 2007; Suto *et al*., 2008; Suto *et al*., 2011).

Findings from our interviews with exam boards

In March 2013, we interviewed senior representatives from seven exam boards – AQA, the CCEA, CIE, Pearson Edexcel, the IB, OCR and WJEC – asking a range of questions about each exam board's arrangements and principles for ensuring marking quality. These interviews referenced the benefits of item-level marking as described above, and also identified a number of other benefits that are not yet supported by any empirical evidence of which we are aware, including:

- Examiners who mark large batches of a particular item mark more reliably than examiners marking whole scripts, because they become very familiar with the mark scheme and a full range of student answers for that question.

- Where an examiner marks all the scripts from a single school or college, there is some evidence that examiners tend to stretch the marks to reflect the full range of the mark scheme. This could mean a school or college with very able students has its marks stretched downwards and a school or college with relatively poorly performing students has its marks stretched upwards. Item-level marking presents examiners with responses from a range of schools or colleges and, therefore, eliminates this effect.

Across the exam boards that do not currently use item-level marking, the IB will be trialling item-level marking across six subjects in 2014, and CIE is considering the introduction of item-level marking for some science subjects. OCR's on-screen marking system allows examiners to mark their scripts by item if they choose, or they can choose to mark at whole-script level. OCR currently has no plans to move to a fully distributed model of item-level marking.

Our approach to this literature review

When we prepared our interim report *Review of Quality of Marking in Exams in A levels, GCSEs and Other Academic Qualifications*, we identified a number of pieces of research that presented a theoretical basis for the benefits of item-level marking over whole-script marking. However, we did not find any empirical research to support these theories. Therefore, in this literature review we focused on empirical studies of the effectiveness of item-level marking in improving marking reliability. We contacted exam boards and asked them to recommend any relevant literature, including grey literature.[3] We only identified a very small number of relevant research papers, which suggests this is a topic yet to be fully explored. We reviewed these papers and the main findings are summarised below.

Findings

Wheadon and Pinot de Moira (2012) looked at marking data from two A level geography units that switched from whole-script marking to item-level marking and then back again to whole-script marking over the course of three years. The study found that the change from whole-script marking to item-level marking appeared to improve the reliability of marking, in particular for the highest performing students.

A study by Black and Curcin (in preparation) compared the reliability of whole-script marking with that of item-level marking. A panel of 12 examiners marked 50 scripts from a unit exam using a counterbalance design. All the examiners marked each script twice: half marked whole scripts followed by item-level marking of the same scripts, and the other half carried out item-level marking before whole-script marking.

The study found evidence of the halo effect at work in the whole-script marking condition, as examiners' perceptions of the overall quality of each student (based upon the student's responses to earlier questions) seemed to affect their awarding of individual item marks to that student. This effect was not present in the item-level marking condition. Item-level marking also seemed to eliminate the most extreme differences between a student's definitive grade and the grade awarded by the

[3] Unpublished studies, studies in progress and studies published outside widely available journals.

whole-script marking. It is noted that although these extreme cases of inconsistency are rare, they do have a significant impact upon public confidence. Therefore, if item-level marking eliminates or reduces the number of cases of extreme examiner error, it could contribute to improved public confidence in the exam system.

However, the study also found strong agreement between examiners. Examiners gave very similar marks to the same students across the whole-script marking condition and the item-level marking condition. Individual examiners also tended not to be consistently lenient or severe across a whole script. If an examiner was lenient or severe, this tended to apply only for specific questions. This finding suggests that there may be little advantage in item-level marking in terms of the impact of severe or lenient examiners.

With regard to the final grade awarded, the study found that item-level marking carried no benefits over whole-script marking in terms of the likelihood of the student receiving the correct grade.

The examiners taking part in the study were also asked what they perceived to be the advantages and disadvantages of the two different modes of marking. The main advantages of item-level marking were identified as speed of marking and greater consistency in applying the mark scheme due to factors such as being more likely to remember the mark given for a previous similar answer. This perception was not supported by the study, which found that the item-level marking condition had a lower Cronbach's alpha[4] (0.752) than the whole-script marking condition (0.795). This suggests that examiners' marking may be slightly more consistent in whole-script marking than item-level marking.

The disadvantages of item-level marking were perceived as boredom leading to possible slips of attention and higher levels of complacency. Whole-script marking was seen as less boring, but examiners thought it took longer to learn the mark scheme under whole-script marking.

Some of the examiners' responses also suggested that the halo effect could be a positive aspect of whole-script marking because:

- whole-script marking may result in a fairer mark overall, as the examiner can be slightly generous on one question (giving the benefit of the doubt) and then balance that with a harsher judgment on another question;

[4] Cronbach's alpha is a measure of the consistency of test scores. This approach splits the test questions into halves and looks at how students perform on each half. This is then repeated for every possible combination of halves, and an average correlation between the halves is calculated.

- examiners gain a greater understanding of the student's ability and overall understanding of the subject;
- examiners gain a better understanding of any transferred errors a student makes, and can adjust their marking accordingly.

The authors suggest that the best mode of marking may be a flexible approach that helps examiners to sustain their concentration throughout the examining period.

Two studies conducted by Pearson Edexcel (2003 and 2004), comparing the reliability of on-screen marking and traditional marking, also provided some indirect evidence of the benefits of item-level marking, although the impact of this research is rather conflated with the impact of online marking.

Pearson Edexcel's 2003 paper reported the findings from an online marking pilot in which papers in GCSE English, GCSE maths and A level maths were marked traditionally and online. In addition, 25 per cent of each maths paper and 100 per cent of each English paper were double marked (marked by two examiners) in both the traditional and on-screen marking format. The online marking was carried out at item level and the traditional marking at whole-script level.

In the maths papers that were marked online, there was no difference between the mean marks, no difference between the standard deviations and a robust correlation between the marks awarded by the first and second examiners. The maths papers that were marked traditionally also showed very high levels of agreement between the two examiners, although the double marking of these papers showed an insignificant difference in the mean marks (0.4 being the most significant difference) and a very small standard deviation (within 0.2) between examiners.

The English papers showed a lower overall correlation between marks across both online marking and traditional marking than seen in the maths papers (as would be expected for this more subjective subject). In traditional marking, the overall mark correlation between the two examiners was 82 per cent for one foundation level paper and 70 per cent for a higher level paper. In online marking, the overall mark correlation between the two examiners across the same papers was 71 per cent for the foundation level paper and 91 per cent for the higher level paper.

This study was repeated in 2004 with a larger sample of GCSE English students across four GCSE English papers. This study also showed high levels of agreement between pairs of examiners in both traditional and online marking. The overall correlation of marks for students where double marking was used traditionally was 80 per cent on the foundation level paper and 76 per cent on the higher level paper. Papers from different exams were double marked using online marking, and the study found the overall correlation of marks for students marked using online marking (at item level) was 94 per cent for the foundation level paper and 92 per cent for the higher level paper.

These findings suggest that online marking at item level is at least as reliable as traditional marking, and possibly more so in some cases. However, it is unclear to what extent item-level marking contributes to the greater reliability of online marking in such cases.

Conclusion

There is currently limited empirical evidence available to enable a robust comparison of the relative merits of whole-script marking and item-level marking. However, the limited research carried out to date suggests that item-level marking seems to be at least as reliable as whole-script marking, and under some conditions is likely to be more reliable than whole-script marking. Specifically, item-level marking may:

- remove the halo effect that means examiners carry forward preconceived ideas about the level of a student's understanding based on his or her answers to previous unrelated items;
- eliminate the most extreme cases of poor marking reliability (which impact significantly upon public confidence), although it is unlikely to make it more likely that a student will receive the correct grade;
- improve the reliability of marking, particularly for the highest performing students.

It is suggested that further research in this area is needed.

References

Black, B. and Curcin, M. (in preparation) *Marking Item by Item Versus Whole Script – What Difference Does it Make?* Cambridge Assessment internal report.

Edexcel, 2003 *The Reliability of Online Marking. Is Online Marking Any More or Less Reliable than Traditional Paper Based Marking? (*Pearson) Edexcel internal report

Edexcel, 2004 *GCSE English Double Marking Project*. (Pearson) Edexcel internal report

Meadows, M. and Billington, L. (2007) *NAA Enhancing the Quality of Marking Project: Final Report for Research on Marker Selection*. London, QCA. Available at: http://archive.teachfind.com/qcda/orderline.qcda.gov.uk/gempdf/184962531X/QCDA104980_marker_selection.pdf (accessed 12th November 2012).

Pinot de Moira, A. (2011) *Why Item Mark? The Advantages and Disadvantages of E-Marking*. Manchester, AQA, Centre for Education Research and Policy. Available at: https://cerp.aqa.org.uk/research-library/why-item-mark-advantages-and-disadvantages-e-marking (Accessed 19th November 2013).

Spear, M. (1996) *The Influence of Halo Effects upon Teachers' Assessments of Written Work*. [no place]. Research in Education, page 85. Cited by Meadows, M. and Billington, L. (2005) *A Review of the Literature on Marking Reliability*. Manchester, AQA. Available at: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf (accessed 12th November 2012).

Suto, I., Crisp, V. and Greatorex, J. (2008) *Investigating the Judgemental Marking Process: an Overview of Our Recent Research*. Research Matters, 5, pp. 6 to 8. Cambridge, Cambridge Assessment. Available at: www.cambridgeassessment.org.uk/ca/digitalAssets/136163_Research_Matters_5_web_.pdf (Accessed 12th November 2012).

Suto, I., Nadas, R. and Bell, J. (2011) *Who Should Mark What? A Study of Factors Affecting Marking Accuracy in a Biology Examination*. Research Papers in Education, 26, 1, pp. 21 to 52. Taylor & Francis Online. Available at: www.tandfonline.com/doi/abs/10.1080/02671520902721837 (accessed 10th February 2014).

Tisi, J., Whitehouse, G., Maughan, S. and Burdett, N. (2013) *A Review of Literature on Marking Reliability Research* (report for Ofqual). Slough, National Foundation for Educational Research. Available at: www.ofqual.gov.uk/files/2013-06-07-nfer-a-review-of-literature-on-marking-reliability.pdf (assessed 10th February 2014).

Wheadon C. and Pinot de Moira A. (2012) *Gains in Marking Reliability from Item-level Marking: Is the Sum of the Parts Better than the Whole?* Educational Research and Evaluation: An International Journal on Theory and Practice Volume 19, Issue 8, 2013. Manchester, AQA, Centre for Education Research and Policy.

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2014

This publication is also available on our website at www.ofqual.gov.uk

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346