# Review of Double Marking Research

February 2014

Ofqual/14/5381

# Contents

# 1. Introduction

This report presents the findings of a review of the possible benefits of double marking over single marking. It covers studies from the 1940s to the present day, focussing predominantly on some of the most comprehensive reviews of double marking research published over the past ten years and supplemented with some more recent, unpublished evidence on the topic. While it considers the most seminal pieces of research on double marking, this report does not attempt to include all the research available on the subject.

In double marking, two examiners independently assess each script (or item). The final mark is the combination of two separate marks. In multiple marking, more than two examiners are used. The combination of double/multiple marks to produce a final score is an acknowledgement that legitimate differences in opinion can exist between examiners, and students may not have one true score. This is fundamentally different from the current hierarchical system used in the UK, in which the marks of the most senior examiner are considered to be the most "true".

While our 2014 report *Review of Marking Internationally* showed double marking is used routinely in some other systems around the world,[1] none of the exam boards offering general qualifications in England currently use double marking in its true sense. Instead, all choose to quality assure marking through a sampling approach.

The assessment systems in those jurisdictions using double marking are generally far smaller in scale than the system in England. The question of whether double marking is truly feasible in England on any scale remains. We discuss the feasibility of double marking in the final section of this report.

[1] www.ofqual.gov.uk/documents/review-of-marking-internationally

# 2. Findings

## 2.1 Evidence for the reliability of double and multiple marking: research findings from the 1940s to 1980s

One of the earliest studies of multiple marking was conducted by Wiseman in 1949 and it considered the composition scripts (essays) of students sitting the 11-plus exam. The independent scores of four examiners were combined to produce a final mark for each script. Wiseman claimed this method produced reliability coefficients of up to 0.946 in this exam, although later researchers have questioned this claim (Wiseman, 1949, cited by Brooks, 2004).

Wiseman's study had two notable features:

1. Individual examiners were not expected to agree with each other. This contradicted the prevailing view, later summarised by Wood and Quinn (1976, cited by Brooks, 2004; p.10) that "disagreement between examiners is reprehensible and… every effort should be directed towards eliminating individual differences." Wiseman (1949, cited by Brooks, 2004; p.8) argued that, provided the examiners were self-consistent, differences in the marking provided a "truer 'all round' picture" of the student's ability.

2. Examiners were trained to use general impression marking; an approach Wiseman claimed could speed up the process so that multiple markings could be achieved through the same time and effort as a single analytic marking.

Wiseman's findings in support of multiple marking were supported by a number of studies carried out in the 1960s and 1970s. Britton, Martin and Rosen (1966, cited by Brooks, 2004) reported an experiment in which 500 O level English language essay scripts were marked by four examiners as well as undergoing the Cambridge Board's official marking procedures. Three of the four examiners marked by general impression and the fourth marked for mechanical accuracy. The results of the experiment showed single marking was "significantly less reliable" than multiple marking (Wiseman, 1949; cited by Brooks, 2004; p.10).

Head (1966, cited by Brooks, 2004) reported on a 1964 experiment in which a sample of essay answers from 290 O level biology scripts was marked by four experienced biology teachers. The average correlation between these teachers was 0.64, rising to 0.84 when the correlation coefficients were calculated from paired scores obtained by averaging the scores of each teacher with every other one in turn.

Another experiment using biology scripts was undertaken by Lucas (1971, cited by Brooks, 2004). This research was notable because it analysed inter-examiner reliability according to whether one, two, three or four separate marks contributed to the final award. Lucas found multiple marking increased the reliability of the marks

awarded significantly, but, more importantly, he found the greatest improvement in reliability came when the number of examiners was increased from one to two. The additional benefits when the number of examiners was increased to three or four were smaller but statistically significant. This position was supported by Kim and Wilson (2009, cited by Tisi *et al.*, 2013), who used data from written compositions and generalizability theory to illustrate that increasing the number of examiners beyond two had little effect on reliability.

Wood and Quinn (1976, cited by Brooks, 2004; p. 15) conducted an experiment in which 100 O level English language scripts were marked by ten examiners using general impression marking. The experiment found the movement from single marking to double marking resulted in a "very considerable reduction in inconsistency". The effect of systematically pairing examiners to take account of known characteristics in their marking was explored, but this delivered a negligible additional benefit.

The main criticism of multiple marking in this period came from Cox (1967, cited by Brooks, 2004; p. 11), who claimed it "does not represent greater agreement on the value of the essays, it is merely a device for getting the same mark every time." However, this objection was "quickly quashed" by Pilliner (1969, cited by Brooks, 2004; p. 11), who demonstrated statistically that Cox's criticism was only valid where examiners were highly self-consistent and strongly disagreed with each other. Provided there was "a fair measure of agreement" across examiners, the aggregated marks would be "a valid expression of the team's consensus", the reliability of which would increase as the size of the team increases.

The concern about possible regression to the mean, resulting in a narrowing of the mark range, was also considered by Wood and Quinn (1976, cited by Brooks, 2004; p. 16). They investigated whether the benefits of greater accuracy of multiple marking outweighed the drawbacks of a reduced spread of marks. Their results showed double marking produced lower variability in marks but it did, nonetheless, "provide better discrimination than single marking." Wood and Quinn also suggested that examiner correlations in the region of 0.50 to 0.60 were sufficient to invalidate Cox's criticism, without being so high as to cancel out the benefits of diversity of opinion between examiners.

These findings encouraged the introduction of double marking in some of the more subjective subjects by a number of exam boards in the 1960s and 1970s. Exam boards also conducted a number of unpublished studies into the reliability of double and multiple marking, such as the Joint Matriculation Board's evaluation of double marking in O level English language and A level general studies scripts (1969, cited by Meadows and Billington, 2005). In O level English language, two papers were marked both analytically and by general impression. The final mark awarded to each student was the total of the two scaled marks. The study found, had only the analytic

marks been used, 6.1 per cent of students on one paper and 6.4 per cent of students on the other paper would have seen their grade change. In A level general studies, two papers were marked twice by general impression and the marks summed to produce the final mark. The study found 6.9 per cent of students would have seen their grade change if only the first impression marks had been used, 7.3 per cent if only the second impression marks had been used.

Most studies of multiple marking added together the marks given to a student by each examiner to form the student's final score. However, Cresswell (1983, cited by Meadows and Billington, 2005) demonstrated that this approach rarely produced the most reliable score possible, and proposed alternative formulae to create a weighted composite score that would optimise the reliability of double marking.

## 2.2 Evidence for the reliability of double and multiple marking: research findings from the 1990s to the present day

Since the 1980s, there have been significantly fewer studies of the reliability of multiple marking in general qualifications. Brooks (2004; p.17) noted that, between the 1970s and the present day, "double marking has all but vanished… as a focus for research activity", as well as from the marking practices of exam boards.

Interest in double marking of general qualifications increased in 2002, when the Qualifications and Curriculum Authority published a report recommending "limited experimental double marking of scripts in subjects such as English" (cited by Brooks, 2004, p. 4) to determine whether double marking could significantly improve marking reliability.

The findings from more recent studies have, on the whole, found less compelling evidence of the benefits of double marking than the studies of the 1940s to 1980s. This, perhaps, reflects the changes that have been made to the nature of assessment of general qualifications since the 1980s.

In a 2005 study, 100 scripts from GCSE English and AS business studies exams were double marked by two groups of examiners, including a senior examiner. This study found a small but significant increase in examiner agreement when paired marks were used. That is, the difference between the senior examiner's mark and the mean of the marks given by a pair of examiners was smaller, on average, than the difference between the senior examiner's mark and the individual examiner's mark (Fearnley, 2005, cited by Meadows and Billington, 2005).

Fearnley compared the level of examiner agreement arising from randomly allocated pairs of examiners to that from pre-selected pairs of examiners and found the increase in examiner agreement was larger for randomly allocated pairs of examiners. This supported Wood and Quinn's findings (1976, cited by Brooks, 2004)

and suggested that there was no additional benefit in systematically pairing examiners in an attempt to balance out extremes of examiner leniency and severity.

However, the random pairing significantly improved the agreement of the marks for only a quarter of the examiners, at best. In addition, the marks of one examiner actually agreed less with the senior examiner's marks after pairing. This led Fearnley to question whether these gains were sufficient to justify the introduction of double marking.

A number of studies have focussed on testing the effectiveness of general impression marking, which proponents argue can enable two or three examiners to mark a script in the same number of person-hours as a single examiner using an analytic mark scheme. This would mean multiple marking not taking longer, costing more or requiring more examiners than single analytic marking (Brooks, 2004). Meadows and Billington (2005) discussed some literature about the pros and cons of general impression (holistic) mark schemes and analytic mark schemes. Tisi *et al.* (2013) cited some more recent studies that suggested general impression marking could be less reliable than analytic marking:

- Blood (2011, cited by Tisi *et al.*, 2013) described a study by Shi (2001) that showed the scores examiners assigned using a holistic scheme did not differ significantly, but examiners did differ greatly in the justifications for their scores. This suggests that, even though the assigned scores were similar, the examiners did not share a common understanding of what it means to be good.

- Ahmed and Pollitt (2011, cited by Tisi *et al.*, 2013, p. 35) argued that a "holistic implicit levels" mark scheme would be less reliable than an analytic levels mark scheme, and the implicitness of the holistic mark scheme was the main source of examiner unreliability.

In contrast, Baker *et al.* (2008, cited by Tisi *et al.*, 2013), in their study of international transferability of National Curriculum Key Stage 3 English marking, found Australian examiners, who were used to holistic mark schemes, expressed concerns about the difficulties of using an analytic mark scheme. Many examiners found the different strands of the mark scheme and their associated multiple criteria disconcerting, and expressed concerns that the marks they had awarded to one strand influenced their marking of other strands.

Cambridge Assessment has recently produced a number of unpublished papers exploring the potential benefits and drawbacks of double marking. In one of these papers, Gray (2010; p.2) argued that, under conventional double marking, the outcome for any script where one or both of the examiners were inconsistent was "entirely unpredictable". (p. 2) In contrast, the re-marking carried out as part of standardisation and sampling assumes the more senior examiner is a more

consistent examiner, and aims to bring the other examiners' marking in line with that of the more senior examiner.

Also for Cambridge Assessment, Bramley (2010; p. 10) held a different view on the accuracy of double marking. Through a mathematical analysis of the possible outcomes for each individual script, he found double marking to be at least as good as single marking, and better in some scenarios. However, he also noted double marking did not improve outcomes over single marking "by as much as might have been anticipated".

Black (in prep) used live marking data from OCR GCSE and A level exams to test the benefits of double marking over single marking. The research drew on examiner scores given to seed scripts in six subjects (geography, English, English literature, economics, critical thinking and psychology), and, as such, the data was gathered under authentic marking conditions rather than a trial. However, the study did not model in the impact of any moderation between examiners, which may form part of double marking. Black's study was unique in terms of its size, using 512,224 marks from 21,562 single marking events. Whereas previous studies had measured marking reliability across whole scripts and/or high tariff essay questions, Black's study used item-level data.

In common with the Cambridge Assessment studies, Black found double marking only delivered small improvements in marking reliability across a range of item types, compared with single marking. These improvements would only yield minimal gains to classification accuracy – the likelihood of students receiving their true grades. However, it is still feasible that double marking might have a more significant impact in a particular subject or unit type. Black's study into such variations is on-going.

## 2.3 The feasibility of double marking of general qualifications

The logistical and financial challenges surrounding the implementation of double marking were recognised by a number of the earlier studies that supported the theoretical argument for the greater reliability of double and multiple marking. For example, Edwards Penfold (1956, cited by Brooks, 2004; p.8) wrote that he doubted the feasibility of double marking when "up to twenty thousand children may be examined in a given year".

Both Brooks (2004) and Meadows and Billington (2005) suggested that double marking of general qualifications may be even less feasible in the present day, due to:

1. Increasing numbers of students sitting more exams: "whereas around 2.5 million O level and A level scripts were processed in the 1970s, the

corresponding figure for GCSE and A level has soared." (Brooks, 2004, p. 5.) In 2012, this figure stood at around 16 million scripts for the summer series alone.[2]

2. Problems with the supply of examiners. "Awarding bodies struggle to recruit enough examiners to mark scripts once, let alone twice." (Meadows and Billington, 2005, p. 58.)

Increased use of on-screen marking in general qualifications, since its introduction in 2003, could make double marking more feasible than when all scripts were marked using pen and paper. Brooks (2004) noted the introduction of on-screen marking had mostly eliminated the logistical problems of transporting the scripts from one examiner to the next, and the additional systems and paperwork required to keep track of this process. Similarly, on-screen marking allows the same script to be independently marked by multiple examiners simultaneously, removing the problem of the additional time it would take to have the script marked by two examiners one after the other.

However, even taking the practical advantages of on-screen marking into account, the resource implications of double marking are substantial. It is also the case that around a third of scripts are still marked traditionally using pen and paper. It would be difficult to justify using double marking only for scripts that are marked on-screen.

The most significant practical barrier to the introduction of double marking is the need for double the number of examiners if it were adopted universally. This may not be possible given all exam boards draw upon the same pool of examiners. Even if recruiting these additional examiners were possible, introducing such a huge number of inexperienced examiners into the system at one time brings obvious risks to quality of marking. The costs of this additional workforce would also be significant. Examiner remuneration is one of the largest costs for all exam boards, and double marking were to be applied universally it would effectively double this cost.

Substantial investment would also be needed to adapt and update exam boards' current computer systems to handle a double marking system. Even to test and pilot double marking across a limited number of subjects would require significant IT system development. These additional workforce and technology costs would need to be covered through increases in the exam entry fees charged by exam boards.

Given that double marking appears to generate only a small increase in the reliability of marking of today's general qualifications, it may not justify the additional costs involved, particularly when other quality control methods may be more cost effective. In his paper on the reliability of marking of GCSE scripts in maths and English,

[2] www.ofqual.gov.uk/standards/research/quality-of-marking

Newton (1996) noted a "trade-off has to be made between reliability and cost-effectiveness: with the very large increase in examination costs that even double marking would incur it would have to yield very much more reliable results than single marking to be considered appropriate" (p. 418).

It has been suggested that, given its additional cost, double marking should be targeted at exams "where genuine benefit can be demonstrated" (Brooks, 2004, p. 21) and not at exams, such as GCSE maths, that already show high levels of inter-examiner reliability (Newton, 1996). However, it is by no means clear which papers at which levels could derive sufficient benefits from double marking to offset the increased costs it would bring (Brooks, 2004). It may also be the case that the subjects where double marking is most likely to bring benefits (usually the more subjective subjects) are the subjects for which exam boards have found it most difficult to recruit new examiners. More research is needed here.

Undoubtedly, there are other means of improving marking reliability at a fraction of the cost of double marking. Research shows question paper and mark scheme design can have the biggest impact on marking accuracy and reliability (Meadows and Billington, 2005). We, therefore, believe the greatest improvements in marking reliability could be delivered by improving the quality of question papers and mark schemes.

Notwithstanding these reservations, there are emerging technological advances that may increase the feasibility of double marking in the future:

- The advent of item-level marking makes it possible to target double marking to the individual items on a paper that are most difficult to mark. Research is on-going to understand exactly what items might benefit from this approach.
- Lamprianou (2004, cited by Meadows and Billington, 2005) suggested that in the future it may be possible to have each script marked by a human examiner and writing assessment software. This would eliminate the additional costs involved in having each script marked by two examiners.

# 3. Conclusion

There is a strong body of evidence from the 1940s to 1980s that double marking is a more reliable method of marking than single marking. However, far fewer studies of double marking have been conducted in the last 20 years, and those studies suggest double marking is only slightly more reliable than single marking.

There are significant logistical and financial challenges associated with the implementation of double marking. If double marking generates only a small increase in the reliability of marking of current general qualifications, it may not justify the additional costs involved, particularly when other quality control methods may be more cost effective.

# 4. References

Black, B. *Double Marking Simulation Study – Interim Findings*. (unpublished). [no place], [no publisher].

Bramley, T. (2010) *Double Marking – How Much Difference Could It Make?* (unpublished). [no place], [no publisher].

Brooks, V. (2004) *Double Marking Revisited*. British Journal of Educational Studies, v52 (n1), pp. 29 to 46. Taylor & Francis, Ltd

Gray, E. (2010) Double Marking: Are There any Benefits? (unpublished). [no place], [no publisher].

Meadows, M. and Billington, L. (2005) *A Review of the Literature on Marking Reliability.* National Assessment Agency . Available at: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf (accessed 3rd February 2014).

Newton, P. E. (1996) *The Reliability of Marking of General Certificate of Secondary Education Scripts: Mathematics and English*. British Educational Research Journal, v22, (n4), pp.405 to 420. Wiley-Blackwell. Available at http://www.tandfonline.com/doi/abs/10.1080/0141192960220403 (accessed 11th 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 (accessed 3rd February 2014).

## Secondary references

### Cited by Brooks (2004)

Britton, J. N., Martin, N. C. and Rosen, H. (1966) *Multiple Marking of English Compositions: an Account of an Experiment*. Schools Council Examinations Bulletin, v12. London, HMSO.

Cox, R. (1967) *Examinations and Higher Education: Survey of the Literature*. London: Society for Research into Higher Education.

Edwards Penfold, D. M. (1956) *Essay Marking Experiments: Shorter and Longer Essays*. British Journal of Educational Psychology, v26 (n2).

Head, J. J. (1966) *Multiple Marking of an Essay Item in Experimental O-Level Nuffield Biology Examinations*. Educational Review, v19 (n1).

Lucas, A. M. (1971) *Multiple Marking of a Matriculation Biology Essay Question*, British Journal of Educational Psychology, v41, (n1).

Pilliner, A. E. G. (1969) *Multiple Marking: Wiseman or Cox?* British Journal of Educational Psychology, v39 (n3).

Wiseman, S. (1949) *The Marking of English Composition in Grammar School Selection*. British Journal of Educational Psychology, v26 (n3).

Wood, R. and Quinn, B. (1976) *Double Impression Marking of English Language Essay and Summary Questions*. Educational Review, v28 (n3).

**Cited by Meadows and Billington (2005)**

Cresswell, M. J. (1983) *Optimum Weighting for Double Marking Procedures*. Associated Examining Board internal Research Unit report no. 281.

Fearnley, A. (2005) *An Investigation of Targeted Double Marking for GCSE and GCE*. London, Qualifications and Curriculum Authority. Available at: http://dera.ioe.ac.uk/9450/1/QCDA104979_an_investigation_of_targeted_double_marking_for_GCSE_and_GCE.pdf (accessed 3rd February 2014).

Joint Matriculation Board (1969) *Report on Double Marking of Essays in General Studies (Advanced) 1969*. JMB Research Report.

Joint Matriculation Board (1969) *Report on Double Marking of Essays in English Language (Ordinary) Papers B and C, 1969*. JMB Research Report.

Lamprianou, J. (2004) *Marking Quality Assurance Procedures: Identifying Good Practice Internationally*. Report prepared for the National Assessment Agency.

**Cited by Tisi *et al.* (2013)**

Ahmed, A. and Pollitt, A. (2011) *Improving Marking Quality through a Taxonomy of Mark Scheme*. Assessment in Education: Principles, Policy & Practice, v18, issue 3, pages 259 to 278. Taylor & Francis Online.

Baker, E., Ayres, P., O'Neil, H. F., Chli, K., Sawyer, W., Sylvester, R. M. and Carroll, B. (2008) *KS3 English Test Marker Study in Australia: Final Report to the National Assessment Agency of England*. Sherman Oaks, CA: University of Southern California.

Blood, I. (2011) *Automated Essay Scoring: a Literature Review.* Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics. Available at: http://journals.tc-library.org/index.php/tesol/article/download/745/470 (accessed 3rd February 2014).

Kim, S. C. and Wilson, M. (2009) *A Comparative Analysis of the Ratings in Performance Assessment Using Generalizability Theory and the Many-facet Rasch Model*. Journal of Applied Measurement, 10: 408 to 423.

Shi, L. (2001). '*Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing', Language Testing, 18, 303–325.* Cited in: Blood, I. (2011). '*Automated essay scoring: a literature review'*. Working Papers in TESOL & Applied Linguistics, 11, 2, 40–64. Available at: http://journals.tc-library.org/index.php/tesol/article/download/745/470 (accessed 11th February 2012)

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