Double Marking Revisited

by Val Brooks

Institute of Education, University of Warwick

Abstract

In 2002, the Qualifications and Curriculum Authority (QCA) published the report of an independent panel of experts into maintaining standards at Advanced Level (A-Level). One of its recommendations was for: 'limited experimental double marking of scripts in subjects such as English to determine whether the strategy would significantly reduce errors of measurement' (p. 24). This recommendation provided the impetus for this paper which reviews the all but forgotten literature on double marking and considers its relevance now.

Key words: double marking, A-Level, essays, inter-rater reliability

1. Introduction

The reliability of essay tests in subjects such as English has long been a subject of concern. Poor levels of inter-rater reliability have been a focus for particular attention. Examining boards have published the results of their own inter-rater reliability studies only infrequently. Nevertheless, several published studies have investigated this topic in an operational context. For instance, Black (1962) investigated: 'the effectiveness of the briefing that assistant examiners receive' (p. 68) by arranging for 19 Ordinary Level (O-Level) examiners to receive a copy of the same English Language script ten days after they had been briefed and whilst they were in the middle of their official marking stint:

> The marks for the essay vary from 54% to 24% ... We can only speculate to what extent briefing had reduced these variations; it had certainly not removed them. (p. 65)

Lucas (1971) also investigated the inter-rater reliability of essay tests under operational conditions, using scripts completed as part of an Australian Matriculation Biology examination. The experiment entailed six examiners assessing the same 44 scripts during their official examination marking. The opportunity for large discrepancies between markers to emerge was reduced in various ways. For instance, a restricted mark range of 0 – 6 was used, with 0 reserved for candidates who failed to answer the question or answered it completely irrelevantly. Notwithstanding these attempts to limit marker variability, the results showed that:

- only one of the 44 scripts had been awarded the same mark by all six markers;
- 19 scripts had a range of two marks;
- 12 had a range of three marks (e. g. awarded 2 by one marker and 5 by one of the others).

Perhaps the most interesting finding is that 12 scripts were awarded a mark of 0 by one marker and 3 by another. Another script was awarded both a 0 and a 4. As Lucas observes: 'Clearly there was not agreement on what constitutes "completely false interpretation of biological concepts" or " complete irrelevance" '(p. 82). A more recent British study (Newton, 1996) compared the reliability of the Southern Examining Group's General Certificate of Secondary Education (GCSE) examinations in Mathematics and English. The reliability of marking in Mathematics was shown to be 'extremely high' whereas that for English was 'notably lower' (p. 405).

Findings such as those described above provided the impetus for the development of double marking, a procedure that entails subjecting each script to two independent assessments. The final award is thus an amalgam of two separate marks. By the middle of the twentieth century, double marking had become the subject of considerable interest. It had been adopted by several examination boards, especially for examinations where the assessment was known to be subjective, for instance English Language and

English Literature, and in newer subjects such as A-Level General Studies. Research by the author in the late 1970s found that a substantial minority of General Certificate of Education (GCE) and Certificate of Secondary Education (CSE) boards was using more than one marker to assess English Language composition scripts completed as part of O-Level or CSE examinations (Brooks, 1980). Six of the eighteen Mode 1 examinations included in the survey employed a system of double or multiple marking. These developments at operational level were stimulated by research that demonstrated the impressive gains in reliability to be made by replacing single marking with double marking.

Between then and now, double marking has all but vanished, both as a focus for research activity and from the practices of the awarding bodies responsible for school-leaving qualifications. It has, however, flourished in higher education where a growing number of university assessment policies require students' coursework to be double marked. But this expansion in use at degree level is set against limited interest in double marking as a research topic. Double marking has, in fact, received little attention at any level in the literature published towards the end of the twentieth century. Exceptions to this generalisation are, predictably, at higher education level (e.g. Partington, 1994; Sparks and Ballantyne, 1997; Smith et al., 2002) and simply underline the point: even where practice is widespread, it has not been matched by research activity. This point is all the more surprising for two reasons. The first is the focus on research in higher education institutions and the second is that inter-rater reliability studies at degree level have produced results that are just as startling as those at school level. For instance, Newstead and Dennis (1994) asked 14 experienced Psychology examiners, all of whom acted as external examiners on other courses, to mark the same six undergraduate scripts. Their marks varied dramatically, the most extreme example involving an essay that received an excellent first from one examiner (mark of 85) and a borderline second/third classification from another (mark of 50).

The decline in interest in double marking in school-leaving qualifications can be explained in part by growing problems with the supply of examiners. In a situation where awarding bodies struggle to recruit enough examiners to mark scripts once, a marking procedure that could double the demand for markers is unlikely to attract attention. It is, therefore, surprising to find that double marking has become the focus of renewed interest. In 2002, QCA published the report of an independent panel of experts into maintaining standards at A-Level. In the section on quality of marking, the report recommended: 'limited experimental double marking of scripts in subjects such as English to determine whether the strategy would significantly reduce errors of measurement' (p. 24). It is this event that provides the background to this paper. The sections which follow revisit the all but forgotten literature on double marking to review key findings on reliability and viability and to consider their relevance now.

Although much has changed since the 1970s when double marking was in its heyday, A-Level resisted real change until much more recently. Introduced in 1951, A-level came to be seen as an academic 'gold standard' with which successive governments were reluctant to tamper: 'a last bastion of real educational worth' (Clarke, 1996, p. 35). No other school-leaving qualification has stood the test of time so well. However, its golden anniversary was the occasion of considerable upheaval as A-Level underwent its first significant revision in fifty years. Traditionally A-Level had been a two-year, linear course culminating in a terminal examination, typically composed of two papers. In 2000, the most far-reaching reform ever attempted re-cast A-Level as a two-phase qualification. Advanced Subsidiary (AS) Level became the compulsory first phase of a new six-unit A-Level, thereby introducing a brand new qualification halfway between GCSE and A-Level. This interim qualification was intended to provide better progression between GCSE and A-Level so that fewer pupils would drop out, finding the gap between the two unacceptably wide, but for those who did abandon their studies, a one-year qualification

would reduce the number who left with nothing to show for their efforts. This re-configuration was achieved by modularising A-Level, a process which entailed breaking it into six separately examined 'chunks' which could be completed at various stages throughout a course of study.

This re-shaping of A-Level has had far-reaching consequences. Perhaps the most important was the destabilisation of the academic gold standard. A key challenge, according to Stobart and Gipps (1997), was: 'how a new "standard", equivalent to the first year of A levels, is to be created so that it is both reliable and comparable across subjects' (p.89). A further complication of this two-phase qualification is that the final award is composed of six modules, three set at a level lower than the full A-Level standard. A final destabilising factor is the opportunity for candidates to re-sit modules. Re-sits are really intended for those who have failed or done significantly worse than expected. However, as only their 'best' marks count towards their award, candidates have nothing to lose by re-sitting papers in the hope of boosting final grades. Each of these factors represented a challenge to the maintenance of A-Level standards.

The reforms have also been beset by logistical difficulties, bringing the entire system to the brink of crisis. By multiplying the number of components that need to be separately examined, and allowing re-sits *during* the course, the reforms have led to a proliferation of examinations. This effect is amply demonstrated by statistics. Whereas around two and a half million O- and A-Level scripts were processed in the 1970s, the corresponding figure for GCSE and A-level has soared to around 24 million now. Even before 2000, examining A-Level was a huge and complex undertaking but the reforms exacerbated rather than alleviating strains that were already apparent in the system. Problems with the supply of examiners, which had been growing over the previous decade, now became acute. The number of candidates to be examined at each sitting made the system almost unmanageable from schools' point of view too. Schools have

struggled with an unprecedented number of timetable clashes and to accommodate the number of candidates they are presenting for examination at each sitting.

The publication of the first set of post-reform results in 2002 was accompanied by allegations that grade boundaries had been manipulated to align standards with established A-level norms. The furore that followed led to a re-grading exercise in which thousands of scripts were re-marked and an enquiry into A-Level grading was launched. Some of the protagonists in this debâcle also became its casualties: the Chairman of QCA and the Secretary of State for Education. However, QCA has retained its new chief executive. Indeed, his own radical reforming agenda was given a fillip by events over the summer of 2002. Revisions to A-Level marking procedures are amongst his top priorities. It is against this background that the relevance of double marking to prevailing circumstances must be considered.

2. Origins of Double Marking

Reliability can be compromised in various ways: the way a test is administered; variations in candidates' performance from one occasion to another; the use of short tests involving too few questions to sample the domain adequately and iron out inter-question variations; the form a test takes and inconsistencies in markers' behaviour, either when marking on the same occasion - inter-rater reliability - or when re-marking material on a separate occasion - mark re-mark reliability. This paper explores the interplay between the latter two factors. Traditionally, the essay has represented a serious threat to examination reliability. This is because essays are extended, free-response items that preclude reliance on a detailed mark scheme which can be determined in advance, applied systematically and without recourse to an examiner's professional judgement. Fixed response tests - and even short, free response items - are inherently more reliable for those very reasons: a precise mark scheme can be specified in advance and the need for individual examiners to exercise discretion in

applying the mark scheme is minimised. Any marking procedure in which examiners exercise professional judgement introduces latitude into the system. Marking behaviour becomes divergent and the reliability of marking is diminished. This problem has been known about for a very long time. As early as the nineteenth century, concerns about essay marking were being expressed (e.g. Edgeworth, 1888).

In the first half of the twentieth century, many attempts were made to improve the inter-rater reliability of essay marking. For instance, in the USA, Quality Rating Scales were developed to provide exemplification material at various levels. Examiners were instructed to finalise their awards by matching scripts against the exemplification material. In England, attempts to improve reliability focused on analytical marking. Analytical marking entailed identifying the criteria that should inform an assessment and assessing each one separately (e.g. Steel and Talman, 1936). The final mark was, therefore, the product of several separate assessments, all made by the same assessor. Unfortunately, none of these initiatives yielded the desired increase in inter-rater reliability.

A breakthrough finally came in the middle of the twentieth century when Wiseman (1949) reported the results of a study based on a radical new approach: multiple marking. Each script was marked independently by teams composed of 4 markers so that the final mark for each script was the sum of four independent assessments. Wiseman used this method to assess the composition scripts of 11-plus candidates. He claimed that it produced reliability coefficients of up to 0.946 - so high that they approached those to be expected from objective tests – although later researchers (e.g. Lucas, 1971) questioned whether Wiseman was actually measuring inter-rater reliability or the mark re-mark consistency of his markers. Two features of this innovation are notable. First, individual markers were not expected to agree with one another. They were selected for involvement in the trial for their high levels of self-consistency. To

meet this criterion, an examiner had to achieve a mark re-mark correlation of .7 or above. Wiseman justified this approach by arguing that self-consistency: 'is the one single measure which is quite clearly a true consistency and one which is closest allied to the normal concept of test reliability' (1949, p. 204) and that:

> provided that markers are experienced teachers, lack of high inter-correlation is desirable, since it points to a diversity of viewpoint in the judgement of complex material i.e. each composition is illuminated by beams from different angles, and the total mark gives a truer "all-round" picture. (1949, p. 206)

Britton, Martin and Rosen (1966) later acknowledged that Wiseman: 'took the revolutionary step of acknowledging the value of differences between markers' (p.10). Second, markers were trained to use general impression marking, making no marks on scripts, an approach which Wiseman claimed could accelerate multiple marking to the point where it offered a viable alternative to single marking completed analytically:

> marking by general impression is a much quicker (and, therefore, cheaper) process than marking analytically. This means that we can treble or quadruple the time and effort on general impression marking without exceeding the time and effort demanded by a single analytic marking. (p. 205)

Even in the middle of the twentieth century, and with the short compositions produced by eleven-year-olds, doubts about the operational feasibility of multiple marking remained despite Wiseman's claims. For instance, Edwards Penfold (1956) argued that:

> the practicability of his [Wiseman's] proposal seems to be in doubt ... [when] up to twenty thousand children may be examined in a given year ... it seems,

> therefore, that the case for the essay will probably be decided on the errors of a single rather than a composite mark. (p. 129)

To this end, Edwards Penfold (1956) decided to look again at whether analytical marking could be made more reliable. Rigorous procedures were developed for his experiment. Markers were involved in constructing the mark scheme with the intention of securing their full agreement with, and understanding of, requirements. This was followed by a period of standardisation and practice marking sessions. Despite this meticulous preparation, Edwards Penfold conceded that the results of his experiment:

> cannot be said to be any advance over the findings of Cast and Finlayson towards achieving an acceptable consistency between the marks of examiners. The variance ratio is still very high. (p. 133)

This was particularly disappointing: ‘in view of the careful standardisation of the scheme’ and further reinforced doubts about the ability of single marking to yield the desired increase in reliability. As an American team who worked on the same problem noted:

> It looked as if the efforts to improve reading reliability had been going in the wrong direction. The solution, it seemed, was in subjecting each paper to the judgement of a number of different readers. (Godshalk et al., 1966, p.4)

Questions also remained about whether multiple marking was applicable to longer scripts written by more mature candidates. This was tested in an experiment carried out in co-operation with the Cambridge Board (Britton, Martin and Rosen, 1966). Britton et al. devised an experiment in which a sample of 500 O-Level English Language essay scripts was marked experimentally by multiple marking teams as well as undergoing the board’s official marking procedure. Each script was independently assessed by four

markers, three marking by general impression and a fourth for mechanical accuracy. Results confirmed that the improvements in reliability achieved in 11-plus examinations were also possible at O-Level:

> marking by individual examiners (*i.e. the official method*) with very careful briefing and elaborate arrangements for moderation was in fact significantly less reliable than a multiple mark. (p. 21)

Like Wiseman (1949), Britton et al. were mindful of the logistical demands of multiple marking. Although they acknowledged that circulating scripts to four markers might make the whole process 'too protracted' in the view of some boards, they emphasised their belief that the demand for examiners would not become inordinate:

> Rapid impression marking is the least tedious kind of marking and those teachers at present deterred by analytical marking and detailed corrections would find impression marking interesting and rewarding. (p. 28)

Britton et al. also contributed to the debate over marker diversity. As Wood and Quinn (1976) later observed: 'Ever since the Hartog, Rhodes and Burt (1936) work appeared, the prevailing view seems to have been that disagreement between examiners is reprehensible and that every effort should be directed towards eliminating individual differences' (p. 233). Britton et al, on the other hand, supported Wiseman, arguing that:

> examiners where they differ, differ in the areas of their most sensitive discrimination and this is the very element in their judgement that we should wish to incorporate into our assessment. (pp. 10-11)

Thus, although doubts remained about the feasibility of multiple marking, it nevertheless appeared to be established as a technique which could significantly improve inter-rater reliability. However, the following year, it came under attack:

> evidence on increasing the reliability by increasing the number of examiners is not very satisfactory … The improvement does not represent greater agreement on the value of the essays, it is merely a device for getting the same mark every time. (Cox, 1967, pp. 7-8)

This challenge to the validity of multiple marking was quickly quashed by Pilliner (1969) who demonstrated statistically that Cox's criticism was valid only in extremely limited circumstances where:

> Each marker is highly self-consistent and if at the same time each agrees poorly with every other… But this correspondence will be largely independent of the real differences which presumably exist in the merits of the essay. Instead, it will express the obstinacy with which each marker maintains his own judgements … an agreement to disagree. If, on the other hand, there is a fair measure of agreement among individual markers about the scripts' merits, the aggregated marks from a team of markers will be a valid expression of the team's consensus of opinion, the reliability of which will increase as the size of the team increases. In this case, Wiseman's procedure is unobjectionable and Cox's criticism is invalid. (p.315)

This theoretical underpinning added weight to the empirical research that had been undertaken at different levels – 11-plus and O-Level. It provided the theoretical justification for multiple marking by demonstrating statistically its capacity to strengthen reliability. Nevertheless, logistical concerns continued to inhibit take-up of this

procedure. It appeared that unless these obstacles could be overcome, multiple marking was unlikely ever to become the procedure of routine choice in an operational setting.

The next landmark in the quest to improve the inter-rater reliability of essay-marking involved a refinement of multiple marking. Research in the 1960s and 1970s developed a streamlined version of multiple marking which came to be known as double marking. These studies were also significant because they widened the scope of the research to encompass another subject: Biology.

3. Double Marking: the Research Evidence

Double marking was trialled as part of the Nuffield Foundation O-Level Biology Project (Head, 1966). The project sought to replace the traditional O-Level, which called 'almost exclusively for factual recall' (p.63), with a new style of examination focused on inquiry, personal experimentation, the weighing of evidence and interpretation of data. This shift from assessing the lower order skill of factual recall, to focus on higher order skills, also threatened the reliability of the examination. Therefore, the project leaders decided to use only test procedures known to have good levels of reliability e.g. multiple choice and short-answer, open-ended items incorporating much information, thus narrowing answers. Essays were not part of the plan but were finally allowed at the insistence of the 100+ teachers collaborating in the project who argued that essays tested skills which could not be tested another way. Concerns about the reliability of essay marking, coupled with doubts about the practicability of multiple marking, led the team to experiment with a modified version of Wiseman's method.

A sample of 290 scripts was selected for experimental marking from the two thousand candidates who took the Nuffield Biology O-Level in 1964. Four experienced Biology teachers carried out the marking, each marking all 290 scripts. This allowed the average correlation between the four examiners to be calculated. At 0.64, this value was below

the minimum threshold of 0.7 suggested by several studies. However, when correlation coefficients for double marking were calculated by pairing each marker with every other, the average value for *r* rose to 0.84. Head concluded that: 'a much more reliable mark was obtained by using this method than by using one examiner's bare marks'. Like its predecessors, this study was mindful of concerns raised by subjecting scripts to more than one assessment: 'The additional burden on the G.C.E. boards of organising two examiners for each script is probably one which could be met; the use of more than two examiners would be a practical impossibility' (p. 71).

Lucas (1971) also investigated double marking using essays written as part of a Biology examination. He wanted to know how well the procedure would stand up to the rigours of operational conditions:

> In Australia the major concern with the reliability of essay marking is felt by Public Examination Boards ... In the case of the Matriculation Biology examination set by the Public Examinations Board of South Australia in 1969, the scripts of 2,265 candidates had to be marked in a period of 14 days. This pressure is aggravated by the need to use as few markers as possible to reduce intermarker variation ... This study was designed to investigate the reliability of the marking under these conditions. (p. 79)

During their official marking, six examiners also took part in an experiment that entailed the same 44 scripts being marked by general impression with awards based on a scale from 0-6. The distinctive contribution of this research is that it analysed inter-rater reliability according to whether 1, 2, 3 or 4 separate marks contributed to the final award. Thus, the relative gains to be made by scaling up from single to double to multiple marking were ascertained.

Lucas found that even under real time examination conditions: ‘Multiple marking has increased the reliability of the marks awarded significantly, and since there was a significant measure of agreement between the markers ... Cox’s objections to this procedure are not valid’ (p. 83). However, his most important finding was that the greatest increase in reliability ‘resulted from an increase from one to two opinions’ (p. 78). Although Pilliner (1969) had demonstrated that reliability increases as the size of the marking team increases, it was Lucas who observed that the greatest improvement came from increasing the size of the marking team from one to two and that any additional benefits to be derived from using teams of three or more were: ‘statistically significant, but of smaller magnitude’ (p. 78). Since manageability is the critical factor in an operational context, it is important to know this.

Lucas did not question the feasibility of double marking. Instead, he enquired whether the cost of multiple marking was justified: ‘whether the costs involved in increasing the number of markers to three or four is justified is not clear. Such a decision needs to be made in the light of information about the effect that the resulting small additional increases in the reliability of marking of individual subjects would have on the reliability of the matriculation examination as a whole. It may be that the money could be better spent‘ (p. 83).

Double marking had now yielded significantly better results than single marking in the assessment of Biology but would it be applicable to other subjects over which concerns remained? In 1976, Wood and Quinn investigated this question in the context of O-Level English Language composition scripts. Although this study was not carried out in an operational setting, there was an attempt to make experimental conditions as close to examination conditions as possible. The experiment was undertaken using scripts written for the London Board’s 1975 O-Level English Language examination and ‘real’ examiners. Before their briefing in analytical marking, the method employed by the

board, ten examiners convened to mark the same 100 essays using general impression marking. The 'experimental' scripts were subsequently re-marked analytically, to avoid candidates' results being influenced by the experiment.

Wood and Quinn emphasised that although reliability can be undermined by two characteristics of markers' behaviour – bias and inconsistency - the real threat to reliability is inconsistency. Bias is relatively easy to correct using routine standardisation procedures. Inconsistency is a measure of how erratic examiners are in deviating from a steady bias and is much more difficult to rectify:

> The reliability of the examination is critically affected by inconsistency among examiners and therefore it is mainly this factor we are aiming to reduce by a system of double marking. (p. 231)

Forty-five pairings were made possible by the combination of any two of the ten examiners. Wood and Quinn found that:

> Naturally, pair-biases vary greatly according to whether like or opposite bias markers are being paired. The most important conclusion is the very considerable reduction in inconsistency from single marking to pair marking; in no case is any pair inconsistency as large as even the smallest individual value ... we are able to conclude straightaway that double marking does indeed lead to greater consistency – fewer errors, better reliability – than single marking (p.238)

Wood and Quinn also explored the effects of pairing examiners systematically to take account of known characteristics in their marking behaviour but found:

> little to choose, as far as inconsistency is concerned, between selecting pairs of examiners either systematically or randomly. The general improvement on single marking by double marking is much greater than any extra improvement which could be gained from a controlled selection and the administrative attractions of random or quasi-random pairing hardly need stressing. (p. 240)

Wood and Quinn also contributed to the debate over marker diversity by asking: 'Just how much disagreement ought to be tolerated'? They proposed that between-marker correlations in the region of 0.50 to 0.60 were acceptable: 'On the principle of wanting some disagreement but not too much' (p. 235). They argued that such correlations are high enough to invalidate Cox's criticism of multiple marking but not so high as to restrict the interplay of the 'multiple dimensions of judgement'.

Finally, Wood and Quinn addressed concerns about regression to the mean that occurs when marks are aggregated. There were fears that this narrowing of the mark range reduced the capacity of double marking to discriminate between different levels of achievement. Wood and Quinn claimed that the real question is: 'whether the improved accuracy of marking outweighs the effect of a reduced spread of marks' (p. 241). They demonstrated that: 'Double marking, despite the lower variability in marks, does provide better discrimination than single marking' (p. 242).

Thus, by the mid-1970s, the benefits of double marking essay scripts were well-established. Pilliner (1969) had demonstrated statistically that the scores it produced were, except in extreme circumstances, 'truer' than those yielded by single marking and that it was not: 'merely a device for getting the same mark every time' (Cox, 1966, pp. 7-8). Lucas (1971) had shown that the real gains to be derived from increasing the number of markers came from the increase from one to two and not from any subsequent increases. Wood and Quinn (1976) had demonstrated that it tackled the real threat to

reliability, marker inconsistency, without undermining the ability of the assessment to differentiate between different levels of performance. A final strength of double marking is that it made a virtue of a feature of marking which had proved impossible to eliminate: diversity of opinion between examiners. Open-ended, free-response test items are highly prized by many subjects because they are seen as the best way of assessing higher order skills. Nevertheless, disagreements about the quality of such responses are always likely to occur. If the professional community agrees that essay-type assessments are valid, then different markers are always likely to see different qualities in the same piece of work. In a single marking system, this is a weakness to be overcome whereas double marking embraces this diversity of judgement, using it to strengthen reliability. A final bonus is that double marking can be completed rapidly, by general impression, dispensing with the need for slower, more laborious methods. Despite its many benefits, between then and now, double marking has all but vanished both from the practices of the awarding bodies responsible for school-leaving qualifications and as a focus for research activity.

This body of evidence provides the context for the independent panel of experts' call for: 'limited experimental double marking of scripts in subjects such as English to determine whether the strategy would significantly reduce errors of measurement' (QCA, 2002, p. 24). Clearly, this evidence is already in the public domain – not specifically at A-level but at several other levels: 11+, 16+ and degree level. It therefore seems reasonable to suppose that A-Level is unlikely to be exempt from these beneficial effects. So perhaps the real question to be considered is not whether double marking is capable of reducing errors in measurement but whether it is viable in the twenty first century when it failed to become established practice in the twentieth?

4. The Case for Double Marking: Some Concluding Observations

Double marking evolved from the attempt to overcome the logistical barriers to multiple marking. Even this leaner variant of multiple marking failed to reassure sceptics. Throughout the period in question (1940s – 1980s) concerns about feasibility remained and only a handful of public examinations ever adopted double or multiple marking. Now, it is hard to find any examples of awards at school level that are double marked. It is important to bear in mind that at the earlier period conditions were comparatively favourable. There were fewer candidates taking fewer examinations and a more plentiful supply of examiners. Even in these relatively favourable conditions, double marking failed to become established as a routine approach to marking subjects where inter-rater reliability was problematic. Two considerations weighed most heavily against the large-scale, long-term use of double marking: cost and the supply of examiners.

Factors which inhibited use in the twentieth century have been greatly exacerbated since that time. One effect of the new unitary A-Level has been a sudden expansion in the number of separately examined components and the number of candidates to be examined on each occasion. Double marking would compound the logistical demands of the reformed A-Level because it is a resource-intensive technique. By doubling the number of assessments to be made, it doubles the demands on other resources, raising implications for every stage in the examination procedure. For instance, the burden on administrative systems would become heavier if scripts need to be forwarded through the post on one extra occasion or copied prior to distribution and there would be additional calculations to compute, enter and double check. If each marker spends the same time per script as in single marking, double marking would require recruitment of examiners on a massive scale or finding extra time for existing examiners. However, the timetable for processing scripts is already tight and offers little room for manoeuvre. Mass recruitment of examiners also looks unrealistic. Alternatively, it may be possible, as Wiseman (1949) claimed, to offset the time burden by each marker spending less

time per script than with single marking. All public examining systems have to juggle the complex and often conflicting demands of validity, reliability **and** manageability. Taken from a philosophical standpoint, the primacy of validity is generally accepted because there is little merit in a test which fails to assess what it purports to assess. In an operational setting, however, manageability is always the ultimate arbiter of use, its demands necessarily taking priority over considerations of validity and reliability. In the current situation, where the proliferation of examinations has brought England to the brink of an examining crisis, double marking offers, perhaps, the least likely way forward.

However, new thinking and innovative practices could alter the prospects for double marking. For instance, the new Chief Executive of QCA has likened examination marking to a cottage industry undertaken at twilight by an army of teachers working at their kitchen tables. He has called for an end to this situation by making examining an integral part of teachers' professional duties to be undertaken during the school day at regional marking centres. If procedures currently being trialled become standard practice, they could lower some of the logistical barriers to double marking to a point where they no longer represent serious obstacles to implementation. Another initiative with the potential to lower logistical hurdles is e-assessment. The ability to take tests online is already available in a limited range of awards. The marking of electronically scanned copies of handwritten scripts is also being piloted. These innovations enable the same script to be marked simultaneously by more than one marker, eliminating postal delays, postal costs and the need to photocopy scripts. The precise location of markers becomes immaterial so examiners working at different ends of the country can assess the same scripts. There is also the potential to lighten the administrative load, and enhance its accuracy, by harnessing the capabilities of information technology. For instance, using on-screen assessment, marks can be entered into electronic mark lists which will compute final scores automatically and forward them to a central database.

Routine use of this kind of innovative practice may be some way away. Nevertheless, when changes to marking arrangements are mooted, a key consideration should be whether they will make double marking more or less viable. This is because a procedure that is capable of enhancing the inter-rater reliability of single marked essay tests, without compromising some other aspect of the assessment, has yet to be developed. For instance, operational calibration has been shown to improve the reliability of single marking (Braun, 1988). This is a statistical technique designed to adjust scores: ‘ to remove the noise contributed by systematic sources of variation; for example, a reader consistently assigning higher grades than the typical reader’ (Braun, 1988, p. 2). It involves embedding a marking experiment within an operational setting. A small, random sub-set of scripts is selected, photocopied and marked by each examiner alongside their normal marking allocation. Statistical techniques are then used to determine the contribution of different sources of systematic variation (e.g. the marker and the stage in the marking period) to the unreliability of the scoring. Raw scores are adjusted accordingly. One concern raised by this approach is that examiners can identify the experimental scripts (photocopied). Indeed, some of Braun’s markers admitted to treating the photocopied scripts differently even though they were instructed to treat the experimental grading as if it were operational. Although some ingenious marking systems have been devised, and the training and standardisation of markers have became ever more sophisticated, the available evidence suggests that the desired increase in reliability is unlikely to derive from any method based on single marking unless the length of each test is greatly expanded. An alternative way of deploying the same extra resources as are required by double marking would be to double the length of each test giving twice as much to mark but with the continuation of single marking. This approach might yield a greater gain in reliability (Ebel, 1972, pp. 250-251) - and possibly in validity too. Although the technical grounds for doubling the size of an examination are persuasive, experience with the introduction of AS Level (2000)

suggests that such a move may be attended by many unwelcome consequences. Thus, double marking remains a firm contender in any attempt to enhance the reliability of A-Level.

Double marking should, however, be carefully targeted at examinations where genuine benefit can be demonstrated. Some examinations already yield high levels of inter-rater reliability - e.g. GCSE Mathematics (Newton,1996) - so double marking would serve little purpose. Whilst it is clear that not all examinations would benefit from double marking, it is by no means clear precisely which papers at which levels could derive sufficient benefit to offset the increased costs of double marking. This is because there is so little up-to-date information in the public domain about the reliability of current awards. Likewise, information about the benefits of double marking is limited to one or two examinations. Fresh research, based on well-designed experimental studies in a range of subjects and types of examination, is needed. Information of this kind is of direct relevance to the body charged with overseeing standards in public awards, making it a useful first step for QCA to take in responding to the recommendation of its panel of experts. As Satterly (1994) observed: 'It is time for the legal insistence on publication of school examination results to be matched by similar openness about reliability in the great body of external assessment on which these comparisons rest' (p. 64).

One remaining obstacle to implementation is whether double marking is philosophically acceptable. Established practice and double marking are built on fundamentally different conceptions of what constitutes a 'true' mark. In the current hierarchical system, the work of assistant examiners is overseen by senior colleagues who report to Chief Examiners whose accumulated wisdom and experience makes them the repository of standards for particular examinations. The system is built on acceptance that marks are 'truer' the higher up the hierarchy the marker is. Double marking rests on a very different view of what constitutes a 'true mark'. Wiseman (1949) postulates that a true mark:

‘would be that given by the pooled judgment of an infinite number of markers’ (p.203). Wood and Quinn (1976) agree, defining a true score as: ‘the average mark for a script awarded by all (our emphasis) the examiners [which] is the best estimate we have of the script’s value free from examiner idiosyncrasies’ (p. 231). This is not simply a change of marking methodology. Rather it represents a fundamental paradigm shift. If the case for double marking can be made, how the system would accommodate such divergence is far from clear. Could two fundamentally different systems of arriving at a final mark co-exist?

REFERENCES

BLACK, E. L. (1962) The Marking of GCE Scripts, *British Journal of Educational Studies,* 11 (1).

BRAUN, H. (1988) Understanding Score Reliability: Experiments in Calibrating Essay Readers, *Journal of Educational Statistics,* 13.

BRITTON, J. N., MARTIN, N. C. and ROSEN, H. (1966) Multiple Marking of English Compositions: an Account of an Experiment, *Schools Council Examinations Bulletin, 12* (London, HMSO).

BROOKS, V. (1980) Improving the Reliability of Essay Marking: a Survey of the Literature with Particular Reference to the English Language Composition, *CSE Research Project Report,* 5 (Leicester, University of Leicester)

CLARKE, J. (1996) Modularising A-Levels in the Social Sciences, *Social Science Teacher,* 25 (3).

COX, R. (1968) *Examinations and Higher Education: Survey of the Literature* (Society for Research into Higher Education)

EBEL, R. L. (1972) Why is a Longer Test Usually a More Reliable Test?, *Educational and Psychological Measurement*, 32.

EDGEWORTH, F. Y. (1888) The Statistics of Examinations, *Journal of the Royal Statistical Society,* 51.

EDWARDS PENFOLD, D. M. (1956) Essay Marking Experiments: Shorter and Longer Essays, *British Journal of Educational Psychology,* 26 (2).

GODSHALK, F. I., SWINEFORD, F. AND COFFMAN, W. E. (1966) The Measurement of Writing Ability, *College Entrance Examination Board Research Monograph*, 6 (College Entrance Examination Board, New York).

HEAD, J. J. (1966) Multiple Marking of an Essay Item in Experimental O-Level Nuffield Biology Examinations, *Educational Review,* 19 (1).

LUCAS, A. M. (1971) Multiple Marking of a Matriculation Biology Essay Question, *British Journal of Educational Psychology,* 41 (1).

NEWSTEAD, S. E. and DENNIS, I. (1994) Examiners Examined: the Reliability of Exam Marking in Psychology, *The Psychologist: Bulletin of the British Psychological Society,* 7.

NEWTON, P (1996) The Reliability of Marking of General Certificate of Secondary Education Scripts: mathematics and English, *British Education Research Journal,* 22.

PARTINGTON, J. (1994) Double-marking Students' Work, *Assessment and Evaluation in Higher Education,* 19 (1).

PILLINER, A. E. G. (1969) Multiple Marking: Wiseman or Cox?, *British Journal of Educational Psychology*, 39 (3).

QUALIFICATIONS AND CURRICULUM AUTHORITY (2002) *Maintaining GCE A Level Standards: The Findings of an Independent Panel of Experts* (London, QCA).

SATTERLY, D. (1994) Quality in External Assessment. In W. HARLEN (Ed.) *Enhancing Quality in Assessment* (London, Paul Chapman).

SMITH, B., SINCLAIR, H., SIMPSON, J., van TEIJLINGEN, E., BOND, C. and TAYLOR, R. (2002) What is the Role of Double-marking? Evidence from an Undergraduate medical Course, *Education for Primary Care*, 13 (4).

SPARKS, R. and BALLANTYNE, R. (1997) Quality Control Methods in large-scale assessment Procedures using 'double-marking' or 'partial double-marking', *Quality Control and Applied Statistics,* 42 (1).

STEEL, J. H. and TALMAN, J. (1936) *The Marking of English Composition* (Nisbet).

STOBART, G. and GIPPS, C. (1997) *Assessment: A Teacher's Guide to the Issues* (London, Hodder and Stoughton).

WISEMAN, S. (1949) The Marking of English Composition in Grammar School Selection, *British Journal of Educational Psychology,* 19 (3).

WOOD, R. and QUINN, B. (1976) Double Impression Marking of English Language Essay and Summary Questions, *Educational Review,* 28 (3).