View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Digital Education Resource Archive

An Investigation of Targeted Double Marking for GCSE and GCE

Austin Fearnley

November 2005

This report was commissioned by the National Assessment Agency

CONTENTS

ACKNOWLEDGEMENTS

The AQA wishes to thank the NAA for sponsoring the research and the NAA Project Board for monitoring the progress of the study.

Ben Jones played an important role in designing the study. Thanks are also due to the sixty-seven examiners who each contributed to this research exercise by marking one hundred scripts in late 2004. Finally, the author acknowledges the help of Jenny England, Martyn Wright and Frances Walsh in assisting with the execution of the work.

Executive Summary

The aim of the current study was to investigate whether a system of double marking could be devised which would improve reliability without requiring more extensive marking by Principal Examiners. The study used two examination components and with examiners some of whom had marked without seeing each others' marks and annotations and some who had marked while seeing others' judgements. The components chosen for study from the Summer 2004 examination were GCSE English Specification A Paper 2 Higher Tier and GCE Business Studies Advanced Supplementary Unit BUS2, People and Operations Management.

The main questions asked in this research were whether or not double marking using annotated or cleaned scripts improved the reliability of marking and by how much. Further, how much improvement in reliability is there where some markers do see other examiners' annotations and where some markers do not see the evidence of any first instance marking? A random allocation of examiners into pairs gave one method of pairing examiners for double marking purposes. A second method was that of targeted pairings based on examiners' previous examiner performances. The effectiveness of each method was tested in the study.

For each component, a sample of thirty-two assistant examiners was selected; sixteen to take part in the Annotated script study and sixteen in the Cleaned script study. One senior examiner, marked clean scripts and one senior examiner marked annotated scripts to produce a 'true' score against which to compare the assistant examiners' marks. The average marks of examiners also provided another indicator of 'true' score. For each component, a stratified sample of 100 scripts was selected.

The main findings of the investigation can be summarised as follows.

- Considerable variation was found between the mean marks awarded by individual examiners even though they had each marked the same 100 scripts.

- There was a statistically significant but small increase (between 1.3 and 1.5 *per cent* of a mark difference) in consistency of double marking over single marking for both subjects in the Cleaned script study where random pairings of examiners were used.

- Where targeted pairings of examiners were used in the Cleaned script study there was also a small but significant increase (1.2 *per cent* of a mark difference) in consistency of double marking over single marking for both subjects.

- Using prior performance criteria of examiners to produce targeted or pre-selected pairs of markers to optimise marking reliability of the double markers was unsuccessful, although small gains in marking reliability were found. Two of the three studies, for each subject, used targeted pairs of examiners but the numbers of examiners benefiting significantly in reliability through double marking in these two studies were fewer than the numbers of examiners benefiting when pairs had been chosen at random in the third study.

- In the Annotated script study with targeted pairs of examiners there was a slight but statistically significant improvement in consistency of double marking over single marking for English (a gain of 0.6 *per cent* of a mark) but there was no gain for Business Studies (a 0.0 *per cent* of a mark difference). Although the English difference was statistically significant and the Business Studies difference was not, the differences in both subjects were small.

- The numbers of examiners benefiting significantly in reliability through double marking in these two studies were fewer when scripts were annotated (benefiting only one examiner across two subjects) than when the scripts were clean. Using annotated scripts may bias the second instance examiners' marks to be closer to those of the first marker and thereby reduce the effectiveness of double marking.

- The levels of gain in reliability through double marking in each of the three studies in each subject were small in terms of marks and may not provide a strong enough incentive to pursue a double marking strategy.

Gains in reliability through double marking were found to be small but, nevertheless, if such a method were to be made operational it may be undermined by procedural difficulties and extra costs. A number of such factors (for example, more examiner fees, more stationery, more scanning equipment or scanning time, more complex 'track and trace' procedures for postal script location monitoring) which would act against double marking were identified. These factors would need to be taken into account if mounting an operational system as they could offset the small gains in marking consistency to be made by double marking.

1. Introduction

In the current GCE and GCSE system of examining, an assistant examiner's mark aims to reflect the standard of the Principal Examiner, and procedures are put in place to train and standardise examiners to attain that standard. Reliability of marking is defined in this context, therefore, as the degree to which assistant examiners' raw marks correspond to the Principal Examiner's standard. Where assistant examiners' raw marks do not reach that standard, their marks are adjusted to be aligned with it. The basis for making such adjustments is, however, narrow since it is impractical for Principal Examiners to mark more than a few scripts from each assistant examiner. Moreover, in a hierarchical marking structure, for example one which includes an intermediate layer of Team Leaders, there are several levels at which unreliability can intrude.

One method of increasing mark reliability may be to have examiners double mark candidates' scripts. There are two possible benefits of double marking. The first benefit may be to give a much needed boost to marker reliability in certain subject areas where marking by some examiners tends to be of relatively low reliability. A second benefit may be the maintenance of a supply of pairs of examiners in periods when examiner recruitment may be difficult by using pairs of examiners whose paired marks were satisfactory but whose single marks were not sufficiently reliable to use on their own.

A literature review of research into marking reliability, including a review of re-marking exercises, undertaken for the NAA by the AQA (Meadows, 2005) has been a useful resource in this study. Brooks (2004) has also recently surveyed double marking research and noted that low reliability of marking, particularly of English essays, has been an issue for a long time, with poor levels of inter-rater reliability. Despite this low reliability, double marking had not apparently been a focus of research since the 1970s except in some notable exceptions.

Wiseman (1949) showed that using multiple markers improved mark reliability. He studied independent re-marking and attempted, and succeeded, in achieving very high mark/re-mark reliability coefficients, greater than 0.9, for the aggregate of four markers of 11-Plus scripts. Wiseman quoted the Spearman-Brown formula as providing a method for calculating the automatic increase in total reliability of n markers over the reliability of a single marker and advised the use of a self-consistency coefficient, consisting of a mark/re-mark correlation coefficient, as a means of examiner selection. Much later, Chaplen (1969) carried out a double impression marking experiment on a university entrance test in English for non-native speakers, where scripts were not annotated during marking, and he found that high mark/re-mark reliability of around 0.90 could be achieved.

The view that attaining high mark/re-mark reliability coefficients was automatically beneficial, however, was questioned by some to be invalid. Cox (1968) queried the merit in seeking very high coefficients through multiple marking. He argued that multiple marking did not represent greater agreement between raters and that, although the average of a very large number of examiners may have high test/re-test consistency, the validity of marking may have been reduced. Pilliner (1969) in a theoretical statistical analysis concluded, however, that Cox's criticisms were valid only in a particular instance where each examiner was highly self-consistent but correlating poorly with the other examiners, in which case the multiple marking constituted an aggregation of disagreements, rather than reflecting inter-rater agreement. Wood and Quinn commented that between marker correlation coefficient levels of between

0.50 and 0.60 were acceptable since one would want some disagreement but not too much. Too little agreement and Cox's criticisms could apply. Too much agreement and there would be no benefit to be obtained by pooling their judgements. Just how much disagreement is optimal for double marking to benefit most is arguable.

The nature of what a 'true' mark should be received some consideration. Wiseman had argued that a 'true' mark would be that given by the composite judgement of an infinite number of markers. Wood & Quinn (1976), agreed by defining the 'true' mark as the average mark awarded by all examiners. This could be viewed as a democratic 'true' mark. At present, the control of standard is with the senior examiners who oversee the work of the assistant examiners and adjust or review examiners' marks to be in line with their standards. Double marking may only be a small step towards a democratic 'true' mark, however, as pairs of marks may be adjusted, reviewed and standardised by senior examiners in the same way that single examiners' marks are adjusted now.

Britton, Martin and Rosen (1966) had experimented on 500 O-level English language essay scripts and they concluded that multiple marking using rapid impressionistic marking gave greater reliability and validity than single marking and was a practicable option. Lucas (1971) investigated multiple marking of Biology essays and found that double marking improved consistency of marking but that diminishing returns were at play for greater numbers of markers such that the additional costs of triple marking might make it impracticable. Brook's review noted the decline in use of double marking since the 1970s in school examinations. The Joint Matriculation Board (Smith (1969a and 1969b) and Griffin (1977)) conducted unpublished research into its operational examinations in A level General Studies and O level English Language, which incorporated double impression marking. The research concluded that double marking should continue but, later, the JMB along with other school examination awarding bodies discontinued the practice.

Double marking has remained absent from school examinations and, as there is also a growing difficulty in recruiting sufficient examiners, it would now appear to be hard to obtain the extra numbers of examiners required for its return. Partington noted, however, in 1994 that multiple marking was then growing in Higher Education, in essay work in the Arts and Law, but argued that double marking was time consuming and unnecessary when using published mark schemes and moderation by external examiners. This argument may not transfer well to school examinations as it would be impracticable to bring the level of moderation in school examinations up to the level possible in Higher Education given the larger numbers of school candidates and teachers involved. Despite difficulties with examiner supply, Baker *et al* (2002) gave fresh impetus to the issue of double marking in a report commissioned by QCA which recommended that limited experimental double marking of scripts be conducted in subjects such as English in order to reduce errors of measurement. Newton (1996) had previously argued that examinations which were already marked consistently by single examiners, such as GCSE Mathematics, would not benefit enough from double marking to offset the extra costs incurred. He also argued that pairing very inconsistent markers would not be effective because regression to the mean and resulting lack of discrimination may undermine the composite marks. Newton concluded that double marking might be most effective for single examiners with intermediate values of mark/re-mark consistency.

In the 1970s heydays of double marking school examinations, Wood and Quinn conducted double marking research in a GCE Ordinary level experimental examination in English and

found that double marking of English essays did lead to improvements in consistency of marking. They argued that the advantages of this increased reliability offset the reduction in spread of marks, and in the consequent reduction in discriminating power, due to regression to the mean. A procedure for selecting pairs of examiners to double mark together which would give optimum marking reliability was also investigated by Wood and Quinn. Some examiners had been systematically paired on the basis of prior examination data on bias (severity or leniency) of marking. Inconsistency of marking was not used in the pairing selection method as only "incomplete impressions by chief examiners" on inconsistency were available to guide the pairing. Wood and Quinn found that random pairings of examiners performed similarly to targeted, or systematic, pairings in attaining improvement in the consistency of marking, on a sample of 100 scripts. They also noted that although systematic pairing can improve bias better than random pairing, bias could be handled satisfactorily for pairs in the same way that they were already handled for single markers. Rather, it was improvement in consistency that was more important to achieve as it was a lack of consistency that was difficult to treat operationally.

The possible influence on the second instance marker of seeing the first marker's marks and annotations on the scripts were investigated by Murphy (1979) in a study on two samples of 100 GCE O level essay scripts. He found that removing marks and annotations of the previous marker made a considerable difference to the re-marking outcomes and he also noted that the only possible doubt about the conclusion was that two sets of scripts, rather than one, had been used for the independent and dependent marking studies. Meadows and Baird (2005) also found that when senior examiners marked photocopied scripts (clean scripts, without any added marking annotations) and live scripts (bearing annotations) for the purpose of standardising assistant examiners' marks, there was a greater discrepancy between marks for the photocopied scripts than for live scripts. This again shows the possible biasing effect on the second marker of seeing the first set of marks.

The aim of the current study bears similarities with that of Wood and Quinn. It enquires whether a system of double marking could be devised which, when used with examination components and examiners with known characteristics, marking with or without seeing each others' marks and annotations, would improve overall reliability without requiring more extensive marking by Principal Examiners. The feasibility and costs of introducing such a method are also considered in this report. Double marking may only produce valuable increases in reliability in question papers where reliability is an issue and for those examiners whose marking is relatively unreliable. It would be inefficient to conduct double marking in subjects where assistant examiners are generally closely aligned to the Senior Examiner (that is, where very highly specified marking criteria are applied), or with examiners who by their previous performance are known to be reliable. GCSE English Specification A Paper 2 Higher Tier and GCE Business Studies Advanced Supplementary Unit BUS2, People and Operations Management, were therefore selected on the basis of estimates of the relative unreliability of their marking for use in this study. The examination specifications, question papers and mark schemes are published on the internet: see the References page for hyperlinks to the AQA website.

The three main questions asked in the study are:

1. In an annotated marking study, where the annotations of the first marker are not removed, are the absolute differences between the mean of the paired examiners'

marks and the senior examiners' marks smaller than between the individual examiners' marks and the senior examiner's marks?

2. Similarly, does double marking using cleaned scripts (where the markers do not see each others' annotations) improve marking reliability, and by how much?

3. Does the application of a criterion-based strategy for pairing examiners improve marking reliability, and if so by how much?

Three further questions were also investigated.

4. How closely do the assistant examiner scores match the definitions of true scores collected in this study?

5. How much unreliability of marking occurs at question level?

6. May the possible gains in reliability be undermined by procedural difficulties and costs of double marking?

2. Design

To choose inconsistently marked components for the study, indices of unreliability for all written examination components were prepared from the Summer 2004 examination information on number of adjustments made to examiners' marks. These data were collected in order. Further data on examiners were collated from the examiners' feedback forms and examiner adjustment records in the selected components in order to select examiners for the study. The feedback forms contained the record of the senior examiners' ratings of the quality of the assistant examiners' marking performances while the examiner adjustment records showed the levels of adjustments made to examiners' marks and whether or not the marking was errant enough to require a review by senior examiners. The most useful information available comprised the average of the absolute value of the adjustment applied to each examiner's marks and, based on this, two low reliability components with large examining panels were selected for the investigation, in GCSE English and GCE Business Studies. The examination specifications, question papers and mark schemes are published on the internet: see the References page for hyperlinks to the AQA website.

For each component, a sample of thirty-two assistant examiners was selected; sixteen to take part in an Annotated script study and sixteen in a Cleaned script study. One senior examiner (denoted CPEX), marked clean scripts and one senior examiner (denoted APEX) marked annotated scripts to produce a 'true' score against which to compare the assistant examiners' marks. The mean marks of examiners also provided an alternative source of 'true' score.

The **first research question** asked was

Q1. In an annotated marking study, where the annotations of the first marker are not removed, are the absolute differences between the mean of the paired examiners' marks and the senior examiners' marks smaller than between the individual examiners' marks and the senior examiner's marks?

Targeted pairings of examiners were used in an attempt to improve marking consistency compared to the marking of the single examiners. To this end, the sixteen examiners in the Annotated study (Table 2.1) were paired in advance of marking according to the available examiner performance criteria, and in this part of the experiment the second examiners saw their partner's script marks and annotations before undertaking their own marking. Examiner A2 saw A1's marks and comments on each script. Examiner A4 saw A3's marks, and so on. Similarly, the first senior examiner (denoted APEX) marked one particular set of annotated scripts, that is the set of scripts previously marked and annotated by both examiners A3 and A4. This allowed an opportunity to find whether or not the senior examiner's marks might have been influenced by seeing the annotations on the scripts made by two assistant examiners.

Table 2.1 Double marking Annotated script study design

	Annotated scripts – targeted pairs																
Pairing	1		2		3		4		5		6		7		8		
Examiner	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	PEx1= APEX
Script																	
1																	
2																	
3																	
4																	
5																	
…																	
100																	

(Where, for example, A1 denotes the first examiner marking in the Annotated study.)

The **second research question** asked was

> *Q2. Does double marking using cleaned scripts (where the markers do not see each others' annotations) improve marking reliability, and by how much?*

The structure of the study here is similar to that in Table 2.1 except in two important respects. First, the examiners were paired at random. The set of random pairings were pairs C1&C2, C3&C4, C5&C6, …, C15&C16. Second, the annotations of any other markers were not on the scripts. The results for paired examiners were again compared to the results for single examiners. The same 100 scripts have been used in both Cleaned script and Annotated script studies, rather than using different sets of scripts, to try to remove a source of possible doubt over the findings.

The **third research question** asked was

> *Q3. Does the application of a criterion-based strategy for pairing examiners improve marking reliability, and if so by how much?*

This investigation involved a re-analysis of the Cleaned script study, where the examiners in that study had all marked independently, but re-pairing the examiners in a criterion based strategy to attempt to maximise marking reliability. The consistency of the results for the double marks were then compared with those of the single marks. The senior marker in the Cleaned script study was denoted CPEX. Research questions one and three differed in one way: the usage of cleaned versus annotated scripts, both involved targeted pairs of

examiners. Questions two and three differed in one way: the usage of random versus targeted pairings, both questions involved cleaned scripts. The Summer 2004 markers in both studies were denoted ORIG, where ORIG is a composite of many markers.

Three other research questions were also investigated.

Q4. How closely do the assistant examiner scores match the definitions of true scores collected in this study?

Q5. How much unreliability of marking occurs at question level?

Q6. May the possible gains in reliability be undermined by procedural difficulties and costs of double marking?

There were six sets of marks, not all independent of each other, which could be used to act as arbiters of assistant examiners' marks, that is, to act as if they were 'true' scores. These variables have been labelled as follows.

- CPEX: is the set of marks of the senior examiner in the Cleaned script study;
- APEX is the set of marks of the senior examiner in the Annotated script study;
- CAVE: is the mean mark across the sixteen Cleaned script study assistant examiners; and
- AAVE: is the mean mark across the sixteen Annotated script study assistant examiners.
- PEX: is the mean mark of the CPEX and the APEX senior examiners;
- AVE: is the mean mark across all 32 assistant examiners;

Main analyses have been carried out for total marks on the scripts but additional analyses have been conducted for individual examination questions. A section on cost and resource factors affecting operational use of double marking concludes the report.

Performance data of two types were collected on individual examiners to inform on optimum reliability of examiner for the two finally selected components, these were the examiners' adjustments and the examiners' errancy or unreliability of marking as recorded during the summer 2004 examination processing. The sizes of each examiner's adjustments and the ranges of marks to which they were applied were available. A senior examiner's rating on a five-point scale of each examiner's standard of marking (lenient, slightly lenient, at right level, slightly severe, severe) were also available. These data were based on a re-marking during the summer examination processing of between fifteen and twenty-five scripts. Scripts marked by relatively unreliable examiners are normally subject to a wide review of marks near the grade boundaries, at the mark review stage after the awarding meeting. Data on unreliability of marking for individual examiners were available in the form of the action to be taken after awarding: whether a wide review takes place or not, and how wide in terms of marks. A senior examiner's rating on a four-point scale of each examiner's standard of marking (consistent, slightly inconsistent, inconsistent, unsatisfactory) which for practical purposes was only a three-point scale as markers of 'unsatisfactory' consistency were sacked. Grossly erratic marking cannot be corrected by an adjustment as one candidate may need a +3 whilst another needs a –3 adjustment. It is usual to avoid making negative adjustments to such examiners so as not to be too harsh to that subset of candidates who need a positive adjustment rather than a negative one. The main criterion in the current research was to pair together markers of similar levels of reliability, and at the same time to

pair markers judged relatively severe with relatively lenient markers so that a net zero adjustment would ideally be attained. When at least one marker of a pair is very reliable, then double marking may be at best unproductive and at worst less reliable than one of the single markers. Pairing reliable examiners with unreliable ones was to be avoided where possible as when at least one marker of a pair is very reliable, then double marking may be suspected to be less reliable than using a single marker. As noted earlier, Wood & Quinn did not use the examiner level of consistency because of the narrowness of the information available. Although prior inconsistency outcomes featured as the main criterion in the pairing of examiners in the current research, there were, however, difficulties encountered in executing this part of the design, particularly in recruitment of markers to the study.

For each component, a stratified sample of 100 scripts from the Summer 2004 examination was selected. Scripts were stratified by mark awarded. Copies of the scripts were made, with all annotations made and the original marking expunged. Important aims in this selection of scripts were not to select scripts which would not copy well and also to avoid major departures from overall mark distribution in the population of examination candidates in the two components. The use of the same 100 scripts in each of the three studies is meant to strengthen comparisons made between the findings in the current study.

Twenty-five examiners for each component received scripts in late autumn 2004. Seventeen of these examiners, one of whom was a senior examiner, were independent markers in the 'cleaned' script study and they returned the scripts as soon as possible after marking. They also completed and returned a mark form by separate cover in case the marked scripts went missing in the post. The other eight markers were first instance markers in the 'annotated' script study who returned marked scripts to be forwarded to the eight second instance markers. In the case of the BUS2 Annotated script study, one examiner dropped out at a late stage thus reducing the number of pairs by one. Difficulties in examiner recruitment to the study had caused numerous replacements to be sought.

3. Results

The main analyses of absolute differences in marks between assistant and senior examiners follow later in this section but, first, preliminary analyses of examiners' raw marks have been carried out, starting with English. Although the distributions of raw marks for all 32 participating English examiners might be expected to be similar, as they all marked the same 100 scripts, they displayed a considerable variation. English examiners' mean raw marks ranged from 40.6 to 32.4, out of a maximum of 54 marks (Appendix A). There were significant differences between examiners, including both single and paired examiners' marks, within the Cleaned script study: (analyses of variance (ANOVA) gave $F(25, 2475) = 57.307$, $p<0.001$) and a similarly significant difference between examiners within the Annotated script study ($F(25, 2475) = 47.923$, $p<0.001$). Although the mean mark awarded on average by Annotated script examiners (AAVE: 36.3) was statistically significantly less than that awarded by Cleaned script examiners (CAVE: 36.6) ($F(1,3069)=5.821$, $p=0.016$) the difference was only 0.3 marks (Table 3.1).

One might expect that examiners who had seen each others' comments might have similar mean marks. Because examiners were paired, each pair can be tested using *post hoc* Tukey analyses to see if their constituent single examiners were statistically different in mean marks.

In the Annotated script study, two of the eight pairs were significantly different from each other. Of the eight pairs of examiners in the Cleaned script study, however, all marking independently of each other, more pairs of examiners (five) had means that were statistically significantly different from each other. This difference in result between Annotated and Cleaned studies may possibly be the effect of the second instance markers in the Annotated study being influenced in standard by knowing the marks on the scripts of the first markers and causing a reduced difference in their mean marks.

Table 3.1 Summary statistics of four independent measures of 'true' marks

	English raw marks				BUS2 raw marks			
	Annotated scripts		Cleaned scripts		Annotated scripts		Cleaned scripts	
	Targeted pairs		Random pairs		Targeted pairs		Random pairs	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Senior Marker								
APEX	33.6	7.5	-	-	23.2	6.7	-	-
CPEX	-	-	35.1	6.3	-	-	22.8	6.8
Examiner means								
AAVE	36.3	5.5	-	-	28.2	7.2	-	-
CAVE	-	-	36.6	5.6	-	-	25.6	7.2

A feature of the design was that the senior examiner ('APEX') in the Annotated script study could see the marks and comments of two individual examiners, A3 and A4, whilst marking which raised the possibility of finding that examiners could be influenced in marking, even though they are experienced markers. The senior English examiner's mean mark (CPEX, 33.6) was not statistically significantly different from the mean marks of examiners A3 (32.4) and A4 (32.5) which were about one mark lower. The independent senior examiner's (CPEX) mean mark (35.1) was 1.5 marks higher than that of APEX, and the average of all examiners' marks was over 36, suggesting that perhaps APEX might have been influenced to award lower marks in English than would have occurred in independent marking.

The English examiners' mark inter-relationships were investigated using product-moment correlation coefficients between raw marks in each study. The correlation coefficients were transformed and then averaged so that it is possible to see how each examiner's marking correlated with the others. It is desirable for average correlation coefficients involving the senior examiners to be high as that would mean that an average marker would be in good agreement with the senior marker's order of merit of scripts. The marks for APEX had one of the highest average correlations (0.83), with other markers in the Annotated study although this was not the case for CPEX (0.71). There was a great deal of variation between examiners in the sizes of the correlation coefficients. One of the coefficients, for Examiner C2, was extremely low (0.48) while those for Examiners A10 and A3 were also low (0.65 and 0.69) (Appendix B).

As noted earlier, the senior marker in the English Annotated study, APEX, saw the marks of Examiners A3 and A4 while he was marking. APEX's marks agreed highly with examiner

A4's order of merit (0.96) and with examiner A3's (0.92) as both coefficients were very high. For other paired examiners, where one examiner had seen the marks of the other examiner in the pair, the coefficients are also relatively high compared with average correlations across all sixteen examiners in the Annotated study. Four of the eight pairs of examiners had very high correlation coefficients between members of the pairs greater than 0.92, and two further pairs had correlation coefficients exceeding 0.87. In the Cleaned study, the original markers in Summer 2004 had a low correlation with the senior examiners' mark (0.72). The correlation between examiner C2 and the senior examiner was the lowest (0.55), indicating exceptional inconsistency.

Turning next to the analyses of examiners' raw marks in the BUS2 studies, there were significant differences between examiners in the Cleaned script study ($F(26,2572) = 43.399$, $p<0.001$). A similar finding was obtained for the Annotated study ($F(22,2178) = 72.505$, $p<0.001$). As in the English analyses, *post hoc* Tukey analyses were used to see if particular examiners were statistically different from one another in their mean marks. One of the seven pairs of Annotated study examiners' mean marks were significantly different from each other. (That is, the two single constituent examiners within the pair were not awarding different mean marks than each other.) On the other hand, of the eight pairs of examiners in the BUS2 Cleaned script study, all marking independently of each other, more (five) pairs of examiners had means that were statistically significantly different from each other. The mean mark of examiners in the Annotated study was significantly different as a group from that of the examiners in the Cleaned study ($F(1,2968) = 388.472$, $p<0.001$).

Post hoc Tukey analyses were used to test for differences in the senior examiner's ('APEX') mean marks in the Annotated study and those of the examiners whose marks he had seen during marking. APEX 's mean mark was not statistically significantly different from the mean marks of examiner pairs A3 and A4. It is possible that the annotations of examiners A3 and A4 may have biased the third instance senior marker who had seen them whilst conducting his own marking. There was only one other examiner pair (A5/A6) with a mean mark not significantly different from the senior marker.

APEX was one of the lower correlating markers in the BUS2 Annotated study (average coefficient = 0.78), although this was not the case for CPEX whose average coefficient was numerically similar at 0.79, but took a middling position with respect to the other Cleaned study examiners. The average correlation coefficients for Examiners A3 and C6 were relatively low (averages 0.72 and 0.73). In the Annotated script study, the correlation coefficients between paired examiners, where one examiner had seen the marks of the other examiner in the pair, were relatively high compared with average correlations across all fourteen examiners in the Annotated study. Four of the eight pairs of examiners had correlation coefficients between members of the pairs greater than 0.94, and two further pairs had very high correlation coefficients exceeding 0.84. Both the levels of marks obtained and the correlation coefficients for paired markers point to a possible influence on the second marker of knowing the first instance markers' marks and annotations which were visible to the second marker in the Annotated study (see Appendix C).

The average mark for each pair of examiners, on each script, was calculated. The desired effect of averaging the marks for a pair of examiners operationally would be to obtain greater reliability of the double marks than of the single marks. English examiner pairs' mean marks ranged from 39.6 to 32.5, a marginally smaller range than for single examiners. BUS2

examiner pairs' mean marks ranged from 32.4 to 22.8, again a marginally smaller range than for single examiners (32.4 to 21.2 marks).

The main three research questions are treated next and these analyses use absolute differences of examiners' marks rather than the raw marks. CPEX, the independent senior examiners' raw marks, were used as a measure of the candidates' 'true' marks. The difference between an examiner's raw mark on a script and CPEX was calculated for all examiners and scripts. The arithmetical sign (+/-) of each of these difference values was then converted to a '+', that is, the absolute values of the differences were calculated. The absolute values were then analysed using Analysis of Variance.

Annotated study: pre-selected pairs of examiners compared to single examiners

Q1. In an annotated marking study, where the annotations of the first marker are not removed, are the absolute differences between the mean of the paired examiners' marks and the senior examiners' marks smaller than between the individual examiners' marks and the senior examiner's marks?

In the English Annotated script study, the pre-selected pairs of examiners were found to have mean absolute differences from the independent senior examiner's marks (CPEX) which were significantly smaller than those differences obtained for the single examiners. The mean absolute difference from CPEX for the eight second instance single markers was 3.8 marks (7.0 *per cent* of the mark allocation of 54) and for pre-selected pairs of markers marking where the second instance marker could see the marks and annotations on the scripts of the first marker the difference was smaller, at 3.4 marks (6.4 *per cent*). That is, the paired marks were closer to the senior examiner's marks than were the single marks. The difference between 3.7 marks and 3.4 marks was significant ($F(1,1485)=9.575$, $p=0.002$). **Although there was a significant difference between the level of agreement of double markers with the independent senior examiner compared with the level of agreement between single markers and that senior examiner, the difference was small (a gain in favour of double marking of 0.4 of a mark out of a maximum of 54).** A *post hoc* Tukey analysis showed that only one of the eight second instance markers (A6) had marked significantly more reliably when acting as one of a pair than when marking singly. A significant difference between examiners within group was also found ($F(14,1485)=9.716$, $p<0.001$) showing that examiners differed amongst themselves in the levels of mean marks they awarded.

In the Business Studies Annotated script study, the mean absolute difference from the independent senior examiner's, CPEX, marks for the seven second instance single BUS2 markers was 6.3 marks (11.5 *per cent* of the mark allocation of 55) and for pre-selected pairs of markers marking where the second instance marker could see the marks and annotations on the scripts of the first marker the difference was the same value, at 6.3 marks. **The mean absolute difference for the single markers was not significantly different from that of the pairs of markers and this non-significant difference in favour of double marking was nil when expressed in marks** ($F(1,1287)=0.037$, $p=0.848$)**.** None of the seven second instance markers had their marks improved significantly, to be closer to the senior examiner's marks, when acting in a pair than when marking singly. The lack of a significant difference between single and pre-selected pairs of markers may have been caused by the second instance examiners' knowledge of the first instance examiners' marks which had biased the

second instance marks to be more like the first instance marks, which could not have occurred if marking had been independent, as it was in the Cleaned script study.

Cleaned study: random pairs of examiners compared to single examiners

Q2. Similarly, does double marking using cleaned scripts (where the markers do not see each others' annotations) improve marking reliability, and by how much?

The randomly paired examiners in the Cleaned script studies in English had a significantly smaller mean absolute mark difference from the senior examiner CPEX's mean mark at 3.1 marks (5.8 *per cent* of the allocation of 54 marks) than did the single examiners, at 3.9 marks (7.3 *per cent*) ($F(1,2277)=53.146$, $p<0.001$). Similarly, in Business Studies, randomly paired examiners had a significantly smaller mean absolute mark difference from the senior examiner CPEX's mean mark at 4.5 marks (8.1 *per cent* of the allocation of 55 marks) than did the single examiners at 5.2 marks (9.4 *per cent*) ($F(1,2277)=27.122$, $p<0.001$). **There was therefore a significant difference between the level of agreement of double markers with the senior examiner compared with the level of agreement between single markers and the senior examiner in both subjects. The difference was a gain in favour of double marking of 0.8 of a mark out of a maximum of 54 in English and a gain of 0.7 of a mark out of a maximum of 55 marks in BUS2. Although statistically significant, the gain in reliability was fairly small, being less than one mark.** A significant difference between examiners within group (including both paired and single examiners) was also found in both subjects (English: $F(22,2277)=14.980$, $p<0.001$ and BUS2: $F(22,2277)=15.476$, $p<0.001$. *Post hoc* Tukey tests were used to show that four of the sixteen English examiners and five of the sixteen BUS2 examiners had marked significantly more reliably when acting in a pair than when marking singly, and one BUS2 examiner had marked less reliably as a pair than when marking as a single marker (5 *per cent* significance level).

Cleaned study: pre-selected pairs of examiners compared to single examiners

Q3. Does the application of a criterion-based strategy for pairing examiners improve marking reliability, and if so by how much?

The mean absolute difference in marks of single English markers from CPEX marks was 3.9 marks (7.3 *per cent* of the mark allocation) and for pre-selected pairs of markers marking independently of one another the difference was smaller, at 3.3 marks (6.1 *per cent*). These differences between single and random pairs of markers were significant at the 0.1 *per cent* level (English: $F(1,2277)=34.835$, $p<0.001$). *Post hoc* Tukey tests showed that three of the sixteen examiners had their marks improved significantly, at the 0.1 *per cent* level, when acting in a pair than when marking singly. (That is, where the averages of pairs of marks more closely matched the marks of CPEX than did their single marks.) The mean absolute difference from the marks of CPEX, for single BUS2 markers was 5.2 marks (9.4 *per cent* of the mark allocation) and for pre-selected pairs of markers marking independently of one another the difference was smaller, at 4.5 marks (8.1 *per cent*). These differences between single and random pairs of markers were significantly different from each other (BUS2: $F(1,2277)=24.678$, $p<0.001$). Two of the sixteen examiners had their marks improved significantly, at the 0.1 *per cent* level and 5 *per cent* level respectively, when acting in a pair than when marking singly. **There was therefore a significant difference between the level of agreement of double markers with the senior examiner compared with the level of agreement between single markers and the senior examiner in both subjects. The**

difference was a gain in favour of double marking of 0.6 of a mark out of a maximum of 54 in English and a gain of 0.6 of a mark out of a maximum of 55 marks in BUS2. The gains, though significant, were not large.

True marks

Q4. How closely do the assistant examiner scores match the definitions of true scores collected in this study?

Significant differences were found between the six variants of 'true' English marks, pointing to no consensus 'true' score being available (English: $F(5,495)=30.647$, $p<0.001$). The mean marks ranged from 33.6 to 36.6. The six variants were in two sets of three. The first three are the two senior markers, CPEX and APEX, and their average, PEX. The second three are the average marks of the two panels, CAVE and AAVE, and the combined average, AVE. Significant differences were also found between the six variants of 'true' BUS2 marks, again pointing to no consensus 'true' score being available (BUS2: $F(5,495)=115.768$, $p<0.001$). The mean marks ranged from 22.8 to 28.2. (See Table 3.1.)

Correlation coefficients between English 'true' marks were relatively high: the smallest being 0.86 (for APEX v CPEX senior examiners) (Table 3.2). Correlation coefficients between BUS2 'true' marks were reasonably high: the lowest BUS2 coefficient was also that between the two senior examiners (0.77) as was also found for English. ORIG, the original marker(s) is included in Table 3.2 for information, and correlations involving ORIG were relatively low which should not perhaps be too surprising for a medley of marks of many different assistant examiners.

Table 3.2 Correlation coefficients between four independent measures of 'true' marks, and the original marker in summer 2004

English	CPEX	APEX	CAVE	AAVE	**BUS2**	CPEX	APEX	CAVE	AAVE
APEX	0.86				APEX	0.77			
CAVE	0.93	0.92			CAVE	0.87	0.86		
AAVE	0.91	0.93	0.98		AAVE	0.88	0.85	0.98	
ORIG	0.72	0.69	0.75	0.74	ORIG	0.73	0.68	0.78	0.79

Correlation coefficients between individual examiners' and the independent senior examiner's marks (Appendices B and C) lie in the range 0.69 to 0.87 in each subject. Correlations between double marks and the senior examiner's marks lie approximately in the top half of that range, from 0.78 to 0.90.

Question Marks

Q5. How much unreliability of marking occurs at question level?

Absolute differences in mean marks between assistant examiners and the senior examiner for English ranged from 1.4 (Q2) to 2.0 (Q1) *per cent* of the mark allocations, for the Cleaned scripts. Values for BUS2 were somewhat less, at 0.5 (Q2) to 1.5 (Q3) *per cent* of the mark allocations. The questions and mark schemes are available on the AQA website (see References).

4. Discussion

Did double marking improve marker reliability?

Were the doubled marks (or more accurately, the average marks of a pair of examiners on a script) closer to the senior examiner's marks than were the marks of the two individual markers? The levels of gain in reliability through double marking found in each of the three studies in each subject, though most were significant, were quite small in terms of marks and may therefore not provide a strong enough incentive to pursue a double marking strategy. The maximum gain in consistency in any of the studies was 1.5 *per cent* of a mark for double marking over single marking. The smallest difference in consistency for double marking over single marking was in fact a zero difference, 0.0 *per cent* of a mark, for Business Studies in the Annotated script study with targeted pairs of examiners, and it was the only result which was not statistically significant.

In GCSE English Paper 2H, grades A to D were each about nine percentage marks wide. For GCE BUS2, most grades were about seven percentage marks wide. For these grade-to-mark equivalences, the highest gain of 1.5 *per cent* of a mark found in the analyses would represent about one sixth of a grade. Although quoting a gain of one sixth of a grade in mark reliability makes the gain seem more attractive than quoting the same effect in terms of marks, the gains found in the study through double marking seem rather small and probably make the procedure not worthwhile to carry out operationally.

Did seeing the marks and annotations of a previous marker influence the marks of a second instance marker?

In each subject, one of the three studies was an Annotated script study. In this study targeted pairs of examiners was the only method of pairing examiners used. The gains in consistency through double marking found in each subject were small, at 0.6 and zero *per cent* of a mark, and the BUS2 gain was not statistically significant. This study gave the smallest percentage gains in reliability found in this investigation. The advantage of double marking should be less if two sets of marks which are being aggregated were very similar. And in a hypothetical extreme case, if the second set of marks were to completely clone the first set, then there could be no double marking effect as the average of the two marks for a script would simply duplicate that of the single marker. This relates to Newton's comment that double marking would not be cost effective if the single marking were highly consistent, such as in Mathematics. The fact that the marks of those examiners who had been paired in the Annotated study design were relatively highly correlated to one another again suggested that the second markers were influenced by seeing the other markers' judgements on the scripts

For almost all of the examiners in the Annotated study there was no gain in reliability to be had from pairing marks. Only one English examiner and no BUS2 examiner in the Annotated script study performed significantly better (that is, matched CPEX's marks more closely) when the average of the pair of marks was used instead of the single marks. The numbers of examiners found in the Cleaned script studies showing significant improvements were about four times greater than this in each subject. The beneficial effects of double making on reliability may be reduced when pairs of examiners' marks are highly correlated. Correlation may be naturally high where examining is highly consistent, for example when assessing to a highly specified mark scheme in mathematics, but perhaps in this case the correlations are artificially high when the second instance marker has a tendency to award marks in line with

the prior marks visible on the scripts. This finding throws doubt on the validity of using annotated scripts in a double marking context.

The finding that the first marker's annotations may bias the second marker when compared with an independent re-marking study is in line with previous research by Meadows and Baird (2005). Murphy, too, found that "Where the original marks and comments had been removed from the scripts, there were much greater variations between the marks which the first senior examiners awarded and the original marks".

Targeted and randomised approaches to the pairing of examiners

Where targeted pairs of examiners had been used in the Annotated script study, there was only a slight gain in consistency of double marking over single marking for English and none in Business Studies. Where targeted pairs of examiners had been used in the Cleaned script study, there were slight gains in consistency but bigger ones than for the Annotated study. The numbers of examiners benefiting significantly in reliability through double marking, however, in these two studies were fewer than the numbers of examiners benefiting in consistency when pairs had been chosen at random in the Cleaned script study. Using prior performance criteria of examiners to produce targeted or pre-selected pairs of markers to optimise marking reliability of the double markers in this study was therefore less successful than the method of random allocation of examiners in pairs. Pairing markers gave a slight improvement on single marking, for both random and targeted methods of pairing, though random pairing is clearly a more efficient and convenient method to conduct than targeted pairing. This finding of the relative merit of random pairings repeats the finding of Wood and Quinn. Unless further work is undertaken on better ways of targeting examiners for pairing for double marking, the outcome of the current study is that a random method would be optimum. A random method would also be more straightforward and efficient to operate. The slightness of the gains in consistency of using any method of pairing examiners may, however, make double marking not worthwhile. Wood and Quinn had also found that double marking led to improved marking reliability.

Did seeing the marks and annotations of a pair of previous markers influence even the marks of a senior marker?

The senior examiner in the Annotated script study, APEX, is the only instance in the study of a third marker acting in a triplet of examiners. In the Annotated scripts study in both subjects the correlations between APEX and the first and second instance markers, A3 and A4, were very high showing a possible influence on even the senior marker of seeing marks and annotations of prior markers on the scripts. The marks of assistant examiners' A3 and A4 in BUS2 also correlated very highly with each other so that senior examiner had seen two highly correlated sets of marks already on the scripts which may have reinforced their effect. The possible influences on the marks of the senior examiners in the Annotated script study confirmed the use of the marks of the senior examiner in the Cleaned script study as the standard of comparison for the analyses carried out of assistant examiners' marks.

What is the nature of a 'true' mark?

Significant differences were found between the mean marks for the six variants of 'true' marks in both subjects, which imply that no consensus 'true' score was available in this study. High correlation between different varieties of 'true' mark is clearly a desirable feature since each

set of 'true' marks should put candidates in the same order of merit. Correlation coefficients between English 'true' marks were relatively high: the smallest being 0.86 (between the senior examiners, APEX and CPEX). The correlation was very high at 0.98 between the 'democratic' variants of 'true' score, that is, between the average assistant English Cleaned script study examiners' marks (CAVE) and the average for the Annotated study marks (AAVE); the same high value of 0.98 was also found for BUS2. The lowest correlation coefficient was 0.77 between the two BUS2 senior examiners.

Higher correlations for democratic 'true' scores than for seniors' 'true' scores indicate that the examiners' average marks were more consistent than the senior examiners' marks, but was this at the cost of discrimination? The standard deviations of marks for English CAVE and AAVE were lower (5.6 and 5.5 marks) than for CPEX and APEX (6.3 and 7.5) which implies that regression to the mean has occurred, which is not unexpected when the marks of sixteen examiners' marks have been averaged. However, this was surprisingly not true for BUS2 where the senior examiners had the lower standard deviations of marks (Table 3.1).

Correlations between double marks and the independent senior examiner's marks lie approximately in the range 0.78 to 0.90. Marks for some individual examiners fall within that range. For some other single markers, however, the correlation falls below that range, down to 0.69 showing that double marking appeared to have improved the consistency of marking of those markers who were the least consistent. Nevertheless, correlation coefficients between 'true' marks and individual assistant examiners' marks were reasonably high except in one case of an examiner where the correlation was exceptionally low, 0.55.

Validity of the double marking

Considerable variation was found between the mean marks awarded by individual examiners even though they had each marked the same 100 scripts. This agrees with other studies that wide variation was often found between different examiners' mean marks when they mark the same scripts: for example Wood and Quinn noted "considerable variation between examiners" in their mean marks. At best, about a quarter of the examiners in the Cleaned script study had marks which were significantly more reliable when paired than when single marks. But not all examiners had gained in reliability when marking in tandem. In fact, one BUS2 examiner had marks which were less reliable when paired. In this case, the marks of the one examiner acting alone would be preferred to those of the pair. The case is a useful illustration that it is not necessarily automatic that a set of double marks will be better, whether significant or not, than both sets of constituent single of marks. It is technically possible for double marking to make matters worse by reducing the reliability of the marks, at least in a few instances.

The correlations between individual examiners in the English Cleaned script study ranged from about 0.45 to 0.85 but, when examiners were paired, the range was higher, being from 0.78 to 0.92. These compared with a generally lower range of coefficients in the Wood and Quinn study of between 0.40 to 0.70 between single impression essay markers and 0.55 to 0.85 for paired markers. Lucas quoted a mean correlation of 0.51 between single markers and 0.71 between paired markers. In Business Studies, the correlations between individual examiners in the Cleaned script study ranged from about 0.50 to 0.90 and between paired examiners the range was from 0.73 to 0.89.

If the correlation were artificially high between paired examiners because of the second instance marker having a bias to award marks in line with the prior marks visible on the

scripts, this may lead to concern over the validity of the paired marks. The effect of the bias, however, is expected not to reduce reliability of the paired marks below the level of the more unreliable of the two single markers.

Targeted pairings were chosen based mainly on known consistency ratings of examiners, measured on a narrow footing using light sampling by the senior examiners during the summer examinations. Pairings matched examiners' levels of consistency, while also pairing 'severe' with 'lenient' biases. Recruitment and retention of examiners to the research study caused some difficulties, however, in practice. Also, very inconsistent examiners tended not to receive negative mark adjustments in operational examinations. Pairing examiners like-with-like in terms of their levels of consistency was a criterion used in order not to adulterate the efforts of a relatively consistent examiner by pairing with a less consistent one. Improving on random pairings of examiners did not occur in this and previous research by Wood and Quinn. Also, the level of gain when using random pairs, although significant, was small in terms of marks.

Is double marking practicable?

Double marking using 'annotated' scripts has been shown to be unfruitful for one subject and only very marginally effective in the other subject. Double marking using 'cleaned' scripts has also been shown to achieve only a small improvement in the consistency of marks over single marking. If it were thought desirable to try to benefit from this small gain in consistency, a number of features of running a double marking system, however, would need to be considered. A main requirement would be that a large number of extra examiners would need to be recruited. Double marking had been in operation at least in a small scale in the 1970s in school examinations, but use fell away to nothing. A particular difficulty in re-introducing double marking would be that those subjects which would most benefit traditionally have a high turnover and shortage of examiners. Also the time taken to mark would double. For paper-script marking, the first marker would need to mark quickly so as to have scripts available to the second instance marker early. Two examiners would need to mark in succession at the same rate as a single examiner. The first examiner would be asked not to annotate the script. In a system where the scripts are scanned into a computer and available to an examiner on a home PC, the time taken for a pair of examiners to mark, either impression or analytic, may not be greater for double marking than single marking as marking could be concurrent. Concurrent double marking would be a big improvement on a postal system but the total examining time would nevertheless be doubled and only half the number of candidates' scripts could be treated in a given time period unless the number of examiners were doubled.

Lamprianou (2004), bearing in mind that resource implications might prevent the utilisation of improvements in reliability of marking through double marking, suggested that each script may be marked by both a human marker and by software. A second human marker being brought in as independent marker and arbiter if the first two markers disagreed. The validity of the marking of essay questions by software is, however, questionable at present.

Using two human markers would present twice the number of mark forms for scanning into the computer if marks were not entered electronically. This would imply twice the operator time and the purchase of extra scanners. In an e-marking process, however, the marks could be sent back electronically rather than on paper, but twice the telephone transmission time would be required. Other items would be doubled or undergo a large increase: number of

standardising meetings, numbers of examiners, senior examiners and team leaders, postage stationery, bags and postage cost; travel costs and catering expenses for meetings; and, computer and operator time.

The system of “track and trace” which is used to enable the board to know where scripts are at any time would be harder to implement, and need more staff on Help Desks, if the number of scripts on the highways were to double from 6 million to 12 million and the route of scripts through the process would be more complex. Other new processes would be required or would become more complex: matching examiners most efficiently at random to act as pairs, aggregating pairs of marks into a single mark per candidate, monitoring the standard of a pair of examiners marking in tandem, and adjusting some examiners’ paired marks. These are some of the foreseeable extra costs, procedures and complexities of operating a double marking system which would offset against the gains in reliability of marking.

Further work

Although the search for targeting gains in marker consistency proved to be elusive, if further work were to be conducted it might focus on investigating other strategies for producing optimum pairings of examiners marking independently.

As examiners seeing the annotations of other examiners while marking are shown to be influenced in the marks they award, it may be possible to turn this drawback in the results into an advantage in the training of examiners, for example, by letting ‘severe’ examiners see scripts marked by ‘lenient’ examiners, and *vice versa*, during marker training. There is no evidence collected from this study, however, on whether the influence lasts beyond the marking of annotated scripts into any subsequent period of independent marking.

REFERENCES

AQA (2004) *Specification, Business Studies Unit BUS2*, *June 2004.* Retrieved 30 September 2004 from http://www.aqa.org.uk/qual/gceasa/bus.html.

AQA (2004) *Question paper, Business Studies Unit BUS2, June 2004.* Retrieved 30 September 2004 from http://www.aqa.org.uk/qual/gceasa/bus_assess.html.

AQA (2004) *Mark scheme, GCE Business Studies Unit BUS2, June 2004.* Retrieved 30 September 2004 from http://www.aqa.org.uk/qual/pdf/AQA-5131-6131-WRE-Jun04.pdf.

AQA (2004) Specification, GCSE English A Tier H, Paper 2, June 2004. Retrieved 30 September 2004 from http://www.aqa.org.uk/qual/gcse/eng_a.html.

AQA (2004) *Question paper (replacement paper) and mark scheme, GCSE English A Tier H, Paper 2, June 2004.* Retrieved 30 September 2004 from http://www.aqa.org.uk/qual/gcse/eng_a_assess.html.

Baker, E., McGaw, B. & Lord Sutherland of Houndwood (2002) *Maintaining GCE A Level standards: The findings of an independent panel of experts*. London: Qualifications and Curriculum Authority.

Britton, J.N., Martin, N.C. & Rosen, H. (1966) *Multiple marking of English compositions: an account of an experiment.* Schools Council Examinations Bulletin, v12. London: HMSO.

Brooks, V. (2004) Double marking revisited, *British Journal of Educational Studies*, v52, n1, pp29-46.

Chaplen, E.F. (1969) The reliability of the essay sub-test in a university entrance test in English for non-native speakers of English. In G.E.Perren & J.L.M.Trim (Eds.), *Applications of Linguistics – papers from the second International Congress of Applied Linguistics*, Cambridge, 1969. Cambridge: Cambridge University Press.

Cresswell, M.J. (1983) *Optimum weighting for double marking procedures*, Associated Examining Board internal Research Unit report no. 281.

Cresswell, M.J. (1985) *A review of borderline reviewing*, Associated Examining Board internal Research Unit report no. 370.

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

Griffin, A. (1977) The *reliability of impression and analytic marking in English essays*, Joint Matriculation Board internal Research Unit report no. 416.

Lamprianou, J. (2004) *Marking quality assurance procedures: identifying good practice internationally.* Report prepared for the National Assessment Agency.

Lucas, A.M. (1971) Multiple marking of a matriculation biology essay question. *British Journal of Educational Psychology,* v41, n1, pp78-84.

Meadows, M.L. (2005) (in draft) A *review of the literature on marking reliability*. AQA and NAA.

Meadows, M.L. and Baird, J. (2005) *What is the right mark? Respecting other examiners' views in a community of practice.* Poster presented at the AEA Europe conference in Dublin, November 2005.

Murphy, R.J.L. (1979) Removing the marks from examination scripts before re-marking them: does it make any difference? *British Journal of Educational Psychology*, v49, pp73-78.

Newton, P (1996) The reliability of marking General Certificate of Secondary Education scripts: Mathematics and English. *British Journal of Educational Research*, v22, n4, pp405-420.

Partington, J. (1994) Double marking students' work. Assess*ment & Evaluation in Higher Education*, v19, n1, pp57-60.

Pilliner, A.E.G (1969) Wiseman or Cox? *British Journal of Educational Psychology,* v39, pp313-315.

Smith, G.A. (1969a) *Report on the double marking of essays in English Language (Ordinary), Papers B and C 1969*, Joint Matriculation Board internal Research Unit report no. 15.

Smith, G.A. (1969b) Report *on the double marking of essays in General Studies (Advanced) 1969*, Joint Matriculation Board internal Research Unit report no. 14.

Wiseman, S. (1949) The marking of English composition in grammar school selection. *British Journal of Educational Psychology*, v26, p172-179.

Wood R. & Quinn, B. (1976) Double impression marking of English language essay and summary questions. *Educational Review*, v28, n3, pp229-246.

APPENDIX A Statistics of examiners' raw marks

SINGLE MARKERS

	English raw marks				BUS2 raw marks			
	Annotated scripts		Cleaned scripts		Annotated scripts		Cleaned scripts	
	Targeted pairs		Random pairs		Targeted pairs		Random pairs	
Marker	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ORIG	37.0	7.4	37.0	7.4	25.4	8.4	25.4	8.4
1	36.4	6.5	35.8	6.2	27.6	7.8	22.5	10.7
2	34.0	5.8	39.7	4.2	29.5	8.2	24.4	7.4
3	32.4	8.2	33.4	5.7	26.4	7.0	28.5	7.0
4	32.5	8.3	37.3	7.8	23.4	6.3	27.3	7.8
5	37.0	4.6	33.1	6.5	25.4	10.5	26.7	8.5
6	40.6	5.9	40.6	4.7	24.0	8.7	30.8	6.5
7	36.2	6.3	33.3	6.1	32.4	7.5	23.0	8.6
8	37.6	7.3	36.4	6.4	31.2	8.5	27.6	8.1
9	35.0	6.9	36.8	6.4	26.1	8.0	22.1	7.4
10	35.7	4.3	36.7	6.8	28.0	8.1	28.0	8.4
11	38.1	6.4	35.7	7.6	32.5	7.7	24.4	8.4
12	38.3	5.9	40.2	6.3	32.4	8.9	21.2	8.2
13	38.7	3.7	34.2	6.6	28.7	7.2	23.3	7.6
14	38.6	4.5	33.2	7.5	27.8	6.6	27.4	6.3
15	35.1	7.3	39.1	6.3	-	-	26.8	10.4
16	34.9	6.7	40.0	5.1	-	-	25.8	7.7
Senior Marker								
APEX	33.6	7.5	-	-	23.2	6.7	-	-
CPEX	-	-	35.1	6.3	-	-	22.8	6.8
Means								
AAVE	36.3	5.5	-	-	28.2	7.2	-	-
CAVE	-	-	36.6	5.6	-	-	25.6	7.2

Continued ….

APPENDIX A … continued

Statistics of examiners' raw marks

PAIRED MARKERS

Pairs of markers	English raw marks				BUS2 raw marks			
	Annotated scripts		Cleaned scripts		Annotated scripts		Cleaned scripts	
	Targeted pairs		Random pairs		Targeted pairs		Random pairs	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1&2	35.2	5.7	37.7	4.7	28.6	7.7	23.4	8.6
3&4	32.5	8.1	35.4	6.1	24.9	6.3	27.9	7.0
5&6	38.8	5.1	36.9	5.2	24.7	9.5	28.8	7.1
7&8	36.9	6.6	34.8	5.8	31.8	7.9	25.3	7.9
9&10	35.4	5.2	36.7	6.2	27.0	7.9	25.1	8.5
11&12	38.2	6.6	37.9	6.5	32.4	8.0	22.8	7.9
13&14	38.6	4.0	33.7	6.7	28.2	6.8	25.4	6.3
15&16	35.0	6.8	39.6	5.2	-	-	26.3	8.5
			Cleaned scripts				Cleaned scripts	
			Targeted pairs				Targeted pairs	
			Mean	SD			Mean	SD
	Pairs of markers				Pairs of markers			
	1&11		35.7	6.6	1&14		24.9	7.8
	2&5		36.4	4.6	2&7		23.7	7.5
	3&13		33.8	5.8	3&8		28.1	6.9
	4&9		37.0	6.6	4&16		26.6	7.4
	6&10		38.7	5.4	6&10		29.4	6.9
	7&14		33.3	6.4	9&11		23.2	7.5
	8&12		38.3	5.7	12&13		22.3	7.5
	15&16		39.6	5.2	14&15		27.1	7.7

APPENDIX B Correlation between examiners' raw marks: English

Marker	English raw marks				
	Annotated scripts			Cleaned scripts	
	Targeted pairs			Random pairs	
	Ave. correl.with other examiners	Correl. with APEX	Correl. with CPEX	Ave. correl. with other exam-iners	Correl. with CPEX
ORIG	-	0.69	0.72	-	0.72
APEX	0.83	-	-	-	-
CPEX	-	-	-	0.71	-
1	0.72	0.81	0.79	0.79	0.84
2	0.74	0.78	0.84	0.48	0.55
3	0.69	0.92	0.71	0.71	0.82
4	0.78	0.96	0.84	0.67	0.77
5	0.75	0.81	0.86	0.73	0.84
6	0.78	0.84	0.88	0.71	0.78
7	0.77	0.81	0.81	0.75	0.85
8	0.74	0.81	0.77	0.68	0.77
9	0.73	0.80	0.81	0.75	0.87
10	0.65	0.66	0.70	0.66	0.77
11	0.83	0.84	0.84	0.75	0.84
12	0.78	0.85	0.85	0.64	0.72
13	0.72	0.71	0.71	0.76	0.87
14	0.76	0.75	0.75	0.73	0.85
15	0.74	0.80	0.81	0.73	0.81
16	0.76	0.83	0.83	0.67	0.79
Pairs of markers					
1&2	0.88	0.86	0.88	0.80	0.79
3&4	0.83	0.95	0.79	0.85	0.86
5&6	0.87	0.86	0.90	0.88	0.87
7&8	0.84	0.83	0.82	0.85	0.86
9&10	0.81	0.79	0.82	0.85	0.87
11&12	0.86	0.85	0.86	0.85	0.84
13&14	0.79	0.74	0.79	0.88	0.90
15&16	0.83	0.83	0.84	0.86	0.86

Pairs of markers		Cleaned scripts	
		Targeted pairs	
1&11		0.95	0.87
2&5		0.90	0.83
3&13		0.94	0.89
4&9		0.93	0.87
6&10		0.90	0.83
7&14		0.95	0.90
8&12		0.91	0.82
15&16		0.83	0.86

APPENDIX C Correlation between examiners' raw marks: Business Studies

Marker	BUS2 raw marks				
	Annotated scripts			Cleaned scripts	
	Targeted pairs			Random pairs	
	Ave. correl.with other examiners	Correl. with APEX	Correl. with CPEX	Ave. correl. with other exam-iners	Correl. with CPEX
ORIG	-	0.68	-	-	0.73
APEX	0.78	-	-	-	-
CPEX	-	-	-	0.79	-
1	0.77	0.74	0.73	0.77	0.74
2	0.83	0.79	0.82	0.83	0.76
3	0.72	0.81	0.69	0.76	0.74
4	0.82	0.88	0.80	0.84	0.80
5	0.81	0.76	0.78	0.84	0.87
6	0.84	0.78	0.79	0.73	0.72
7	0.79	0.76	0.74	0.81	0.83
8	0.82	0.79	0.79	0.77	0.75
9	0.82	0.74	0.82	0.83	0.78
10	0.85	0.80	0.83	0.77	0.75
11	0.81	0.78	0.81	0.78	0.79
12	0.77	0.74	0.81	0.79	0.78
13	0.79	0.73	0.75	0.76	0.76
14	0.78	0.72	0.75	0.76	0.80
15	-	-	-	0.81	0.77
16	-	-	-	0.77	0.74
Pairs of markers					
1&2	0.86	0.79	0.80	0.86	0.80
3&4	0.84	0.89	0.79	0.89	0.82
5&6	0.86	0.78	0.80	0.87	0.85
7&8	0.84	0.79	0.78	0.86	0.84
9&10	0.87	0.78	0.84	0.87	0.80
11&12	0.85	0.79	0.85	0.86	0.83
13&14	0.81	0.73	0.76	0.87	0.86
15&16	-	-	-	0.87	0.80

Pairs of markers				Cleaned scripts	
				Targeted pairs	
1&14	-	-	-	0.94	0.83
2&7	-	-	-	0.97	0.84
3&8	-	-	-	0.94	0.81
4&16	-	-	-	0.95	0.80
6&10	-	-	-	0.91	0.79
9&11	-	-	-	0.95	0.83
12&13	-	-	-	0.91	0.82
14&15	-	-	-	0.95	0.84