they conceal. This last point is well illustrated by Vidal Rodeiro (2007, this issue) – the reader is encouraged to compare in her article tables 4 and 11 with tables 5, 6 and 12.

References

Altman, D.G. & Bland, J.M. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician*, **32**, 307–317.

Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307–310.

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213–220.

Cronbach, L.J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability of scores and profiles*. New York: Wiley & Sons.

Eye, A. von & Mun, E.Y. (2005). *Analyzing rater agreement: manifest variable methods*. Mawhah, NJ.: Lawrence Erlbaum Associates.

Harvill, L.M. (1991). An NCME Instructional Module on Standard Error of Measurement. *Educational Measurement: Issues and Practice*, **10**, 2, 33–41.

Krippendorff (2002). Computing Krippendorff's Alpha-Reliability. http://www.asc.upenn.edu/usr/krippendorff/webreliability2.pdf. Accessed January 2006.

Linacre, J.M. (1994). *Many-Facet Rasch Measurement*. 2nd Edition. Chicago: MESA Press

Lord, F.M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Myford, C.M. & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement*, **4**, 4, 386–422.

Myford, C.M. & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part 2. *Journal of Applied Measurement*, **5**, 2, 189–227.

Newton, P.E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal*, **31**, 4, 419–442.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, **86**, 2, 420–428.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis. an introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Stemler, S.E. (2004). A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, **9**, 4. Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4 November, 2005.

Sukkarieh, J.Z., Pulman, S. G. & Raikes, N. (2003). *Auto-marking: using computational linguistics to score short, free text responses*. Paper presented at the 29th conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Uebersax, J. (2002a). Kappa coefficients. http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm Accessed 22/01/07.

Uebersax, J. (2002b). Raw agreement indices. http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm Accessed 22/01/07.

Uebersax, J. (2003). Intraclass Correlation and Related Methods http://ourworld.compuserve.com/homepages/jsuebersax/icc.htm Accessed 22/01/07.

Vidal Rodeiro, C. L. (2007). Agreement between outcomes from different double marking models. *Research Matters: A Cambridge Assessment Publication*, **4**, 28–34.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, **103**, 3, 374–378.

ASSURING QUALITY IN ASSESSMENT

Agreement between outcomes from different double marking models

Carmen L. Vidal Rodeiro Research Division

Introduction

The practice of arranging for students' work to be marked by more than one person is a subject of great interest in educational research (see, for example, Cannings *et al*. 2005, Brooks, 2004, White, 2001 or Partington, 1994). However, deciding if double marking is worthwhile incorporates a perennial dilemma. Intuitively, it seems to increase the reliability of the assessment and shows fairness in marking, but this needs to be proven a benefit in order to justify the additional time and effort that it takes. Awarding bodies struggle to recruit enough examiners to mark scripts once, never mind twice, and therefore double marking of all examination papers can be a difficult task.

In the context of GCSE or GCE examinations, double marking can be a means to enhance the reliability of the marking process. One of the principal concerns of any examination board is to ensure that its examinations are marked reliably. It is essential that each examiner is applying the same standard from one script to the next and that each examiner is marking to the same standard as every other examiner. Although Pilliner (1969) had demonstrated that reliability increases as the size of the marking team increases, it was Lucas (1971) who observed that the greatest improvement came from increasing the size of the marking team from one to two and that additional benefits derived from using teams of three or more markers were of smaller magnitude.

Double marking models

Double marking is more common in examinations where the assessment is known to be subjective, for example, examinations involving writing an essay. In these cases, the main methods of double marking are:

a. **Blind double marking.** The first marker makes no annotations on the work being marked and the second marker examines all pieces of work as left by students.

b. **Non-blind or annotated double marking.** In this case, the first marker makes annotations on the work being marked and the second marker marks it with this information known. This may involve varying degrees of information available to the second marker, for example, annotations to draw attention to points in the text or marks written on answers.

Whatever method is used for double marking examinations, there must be a method of resolving differences between markers. Some of the methods that can be employed for this task are:

a. Discuss and negotiate the marks on all the differences or on specified differences.

b. Take the mean of the marks. This may be done for all differences or for specified differences. However, there are studies that suggest that taking the average of two marks is not the best way to reconcile the differences. For example, Massey and Foulkes (1994) suggested that the average of two blind marks may not always be a sound estimate. It remains at least arguable that the greater the difference between two markers the more likely it is that one has seen something the other has not.

c. Resort to a third marker, who could mark the script afresh or, based on the two previous marks, produce a final mark.

Aim of the research

The main purpose of this research is to evaluate the agreement between marks from different double marking models, in particular, blind and annotated double marking. We focus on agreement concerning total marks across questions in the examination paper (or component) concerned. We acknowledge that future technologies may change the current marking practice so that instead of one examiner marking the whole of a candidate's paper, questions might be allocated individually to different examiners.

Specific aims are:

1. To measure marking outcomes and agreement between first and second marking.
2. To compare second marking workload in relation to the double marking models, including the impact of examiner experience.
3. To measure reconciliation workload (number required plus time taken).

Data and methods

Description of the data and the task

Two General Certificate of Secondary Education (GCSE) units, OCR English and OCR Classical Greek, were selected for this study. For English, one component was chosen: Literary Heritage and Imaginative Writing, Higher Tier. The total number of marks for this unit was 40. For Classical Greek, the component 2, Verse Literature, was selected. The total number of marks for this unit was 60. For each subject, a two hundred script sample from the June 2004 examination was retained.

Five examiners per subject were invited to participate in this research: a principal examiner (PE), two senior examiners (or experienced assistant examiners) and two assistant examiners.

For both English and Classical Greek, the scripts were split into two packs of one hundred scripts. Each assistant examiner was allocated one hundred scripts from a range of different marks. These scripts had all marks and marking annotations removed. Each of the more experienced or senior examiners was allocated two packs of scripts. One pack contained one hundred scripts that had the marks and comments from the original examiners on them, whereas for the one hundred scripts in the other pack, these were removed. In each pack the scripts were from a range of different marks. We ensured that each script appeared in only one pack.

For each subject, the examiners were asked to mark the scripts following the same marking instructions that had been used in the original marking of the examination. A meeting with the examiners took place before the re-marking started. In the meeting, the principal examiners reviewed the mark scheme with the assistant and senior examiners in order to identify any marking issues. It should be noted that this meeting was not a full standardisation meeting and that, as this research was done under experimental conditions, some of the quality assurance procedures that are carried out during live marking were not performed.

Reconciliation was carried out when the difference between the original 'live' mark and the mark awarded in this study for the same script exceeded 10% of the mark range. The principal examiners in each subject performed this task, producing a final mark.

After the marking and the reconciliation were performed, the experiment produced four marking outcomes in each subject:

1. Original: 200 scripts with the original marks awarded in the June 2004 session.

Plus re-marking of the same 200 scripts using three different strategies:

2. Treatment 1: Blind re-marking by two assistant examiners (marking 100 scripts each) plus the reconciliation by the PE as needed.
3. Treatment 2: Blind re-marking by two senior (or experienced) examiners (marking 100 scripts each) plus the reconciliation by the PE as needed.
4. Treatment 3: Non-blind or annotated re-marking by two senior (or experienced) examiners (marking 100 scripts each) plus the reconciliation by the PE as needed.

Statistical methodology

There is little consensus about what statistical methods are best to analyse markers' agreement. There are many alternatives in the literature although the most commonly used are the correlation coefficients and the Kappa statistics (see Uebersax, 2003, for an overview of the different statistics that are used in this field and Bramley, 2007, for a discussion of how they might be applied in a double marking context).

Correlation coefficients

Usually, the first step in this type of analysis is to plot the data and draw the line of equality on which all points would lie if the two markers gave exactly the same mark every time. The second step is to calculate the correlation coefficient between the two markers (ρ) which measures the degree to which two variables are linearly related. When the relationship between the two variables is nonlinear or when outliers are present, the correlation coefficient incorrectly estimates the strength of the relationship. Plotting the data before computing a correlation coefficient enables the verification of a linear relationship and the identification of potential outliers.

On the principle of allowing for some disagreement but not too much, in the context of double marking examinations Wood and Quinn (1976) proposed that between-marker correlations in the region of 0.50 to 0.60 would seem to be realistic.

Measures of agreement

Early approaches to the study of markers' agreement focussed on the observed proportion of agreement, that is, the proportion of cases in which the markers agreed. However, this statistic does not allow for the fact that a certain amount of agreement can be expected on the basis of chance alone. A chance-corrected measure of agreement, introduced by Cohen (1960), has come to be known as Kappa. For two markers, it is calculated as follows:

$$\kappa = \frac{P_a - P_c}{1 - P_c},$$

where P_a is the proportion of marks in which the markers agree and P_c is the proportion of marks for which agreement is expected by chance.

Table 1 shows the degree of agreement for different values of Kappa (Landis and Koch, 1977). The limits of this classification are arbitrary and can vary according to the study carried out. Kappa can take negative values if the markers agree at less than chance level and it can be zero if there is no agreement greater or lesser than chance.

Table 1 : Degree of agreement and values of Kappa

Degree of agreement	*Kappa*
Excellent	≥ 0.81
Good	0.80 – 0.61
Moderate	0.60 – 0.41
Poor	0.40 – 0.21
Bad	0.20 – 0.00
Very bad	< 0.00

Results

Examiners were able to mark, on average, 5 or 6 scripts per hour. This did not seem to vary whether the scripts were annotated or blind. Some examiners originally thought that marking the annotated ones would be swifter but this proved not to be the case. There seems to be no difference between the time employed by assistant and senior examiners in marking their scripts.

GCSE Classical Greek

Table 2 displays summary statistics of the marks awarded in the different marking treatments. The means do not differ very much and the standard deviations are very similar in all cases. The marks given to the scripts are all rather high (the minimum available mark for the component is 0 and the lowest mark awarded by an examiner is 17). The re-markers appear very similar in their overall marks but all mark, on average, more generously than the original markers.

Table 2 : Summary statistics of the marks awarded in the different marking treatments

	N	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
Original	200	43.72	9.06	17	60
Treatment 1	200	44.05	8.82	17	59
Treatment 2	200	43.93	9.15	15	60
Treatment 3	200	44.09	9.19	18	60

Table 3 displays the absolute (unsigned) differences between the original marks and the three sets of re-marks. The average mark change between the original and the first treatment (blind re-mark by assistant examiners) is bigger than for the other treatments. The smallest value corresponds to the non-blind re-mark (treatment 3). This last difference is probably caused by seeing the actual marks awarded but it might, in part, be due to comments providing additional insight into why the original examiner awarded a particular mark.

Table 3 : Absolute differences in marks

	Mean Difference	*Standard Deviation*
Original – Treatment 1	2.16	1.69
Original – Treatment 2	1.97	1.73
Original – Treatment 3	0.67	0.84

The simplest way to describe agreement would be to show the proportion of times two markers of the same scripts agree, or the proportion of times two markers agree on specific categories. Table 4 displays these proportions.

The percentages of exact agreement between the original marks and the different sets of re-marks are 16%, 17% and 50%. When agreement is widened to include adjacent marks, agreement levels increase. For example, for treatment 1 (blind re-marking by assistant examiners) the marks differ by no more than +/- one in around 43% of the scripts marked and by +/- three marks in around 78% of the scripts. For treatment 2 (blind re-marking by senior or more experienced assistant examiners) the marks differ by +/- one in around 45% of the scripts marked and by +/- three in around 87% of the scripts. For treatment 3 (non-blind re-marking) the marks differ by +/- one mark in around 87% of the scripts marked and in three or fewer marks in around 98% of the scripts.

Table 4 : Distribution of differences between original and experimental marks

Difference in marks	*Treatment 1 (%)*	*Treatment 2 (%)*	*Treatment 3 (%)*	*Total (%)*
≤ -6	3.5	1.0	0.0	1.5
-5	2.0	2.5	0.0	1.5
-4	7.5	4.0	1.5	4.3
-3	10.5	9.0	2.0	7.2
-2	10.0	16.0	7.0	11.0
-1	11.5	14.5	26.5	17.5
0	16.0	17.0	50.0	27.7
1	15.0	13.0	11.0	13.0
2	10.5	9.5	2.0	7.3
3	4.5	8.0	0.0	4.2
4	6.0	2.5	0.0	2.8
5	1.0	1.5	0.0	0.8
≥ 6	2.0	1.5	0.0	1.2

Table 4 provides, again, evidence that removing previous marks and comments from scripts does make a difference. There are alternative interpretations of this. A negative perspective would suggest that examiners who are asked to re-mark scripts cannot help but be influenced by the previous judgements, however much they try to ignore them and form their own opinion. A positive view would argue that the non-blind re-markers can see why the original mark was awarded and are happy to concur; even though had they marked blind they might well not have spotted features noted by the original examiner.

Pearson's correlation coefficients between marking treatments are displayed in Table 5.

Table 5 : Pearson's correlation coefficients

	Original	*Treatment 1*	*Treatment 2*	*Treatment 3*
Original	1.0000	0.9538	0.9588	0.9940
Treatment 1	0.9538	1.0000	0.9478	0.9554
Treatment 2	0.9588	0.9478	1.0000	0.9639
Treatment 3	0.9940	0.9554	0.9639	1.0000

The correlation coefficients are high (the smallest correlation appears between the original mark and treatment 1: $\rho = 0.9538$) and of an order which would normally be regarded as an indicator of high reliability of marking. The highest correlation appears between the original marks and the non-blind re-marks. The correlation between the treatment 2 (blind re-mark by senior or more experienced assistant examiners) and the original marks is higher than the correlation between treatment 1 and the original marks, which might reflect the relative experience of the examiners.

Another way of assessing the agreement between pairs of markers is the use of Kappa (Kappa statistics are displayed in Table 6). Again, this table provides confirmation of the hypothesis that the marking of two examiners would be affected by whether or not previous marks and comments had been removed from the scripts.

Table 6 : Kappa statistics[1]

	Original
Treatment 1	0.7609
Treatment 2	0.8103
Treatment 3	0.9327

1 The values of the Kappa statistic provided in this table were not obtained using the formula given in this article but using an extended version (Cohen, 1968).

Reconciliation

Using the 10% criterion described in the methodology section, we determined which scripts needed reconciliation. For Classical Greek the maximum and minimum marks are 60 and 0, respectively. Then, if for a particular script, the absolute difference between two marks is bigger than 6, the script needs reconciliation and this is undertaken by the principal examiner. Table 7 displays the numbers and percentage (in brackets) of scripts that needed reconciliation.

Table 7 : Scripts that needed reconciliation

		Examiner Experience		*Marking*	
	Total	*Blind Assistant*	*Blind Senior/ Experienced*	*Blind*	*Non-blind*
Reconciliation	16 (2.7)	11 (5.5)	5 (2.5)	16 (4.0)	0 (0.0)

Only 16 of the re-marked scripts needed reconciliation (2.7%). Of those, 11 scripts were blind re-marked by assistant examiners and 5 by senior or more experienced assistant examiners. This confirms that when experienced examiners re-marked scripts the differences with the original marks are smaller than when assistant examiners did so. Non-blind re-marked scripts did not need reconciliation.

Only 4 of the reconciliation outcomes correspond with the mean of two prior markings, although 12 of the reconciliation outcomes are within +/- two marks of this mean. Reconciliation for blind marks by assistant examiners produces outcomes more widely distributed around the mean of prior marking than for senior or more experienced assistant examiners (see Table 8). Note that the numbers involved in the reconciliation task are too small to draw any strong conclusions.

Table 8 : Difference between the mean of two marks and the reconciliation outcome

	Treatment 1		*Treatment 2*	
Difference	*Frequency*	*%*	*Frequency*	*%*
-3	1	9.1	0	0.0
-2	0	0.0	1	20.0
-1	1	9.1	2	40.0
0	2	18.2	2	40.0
1	1	9.1	0	0.0
2	3	27.3	0	0.0
3	2	18.2	0	0.0
4	1	9.1	0	0.0

Reconciling differences is likely to prove better than averaging because it takes better advantage of the information available or even gathers and uses some more. However, this approach might be difficult to transfer to large scale public examinations. The fact that non-blind re-marking required no reconciliation may well be an important advantage in large scale operations.

During the reconciliation task, the principal examiner 'reconciled' around five scripts per hour. If we had changed the cut-off point for

reconciliation and reconciled scripts where the absolute difference between two marks was bigger than 3 (5% of the mark range) then the time employed and the cost that it entailed would have made the reconciliation task much more expensive. The total percentage of scripts needing reconciliation would have been around 12%. 17.5% of the blind re-marked scripts and 1.5% of the non-blind re-marked scripts would have had to be reconciled.

GCSE English scripts

Table 9 displays summary statistics of the marks awarded in the different marking treatments. The mean is half a mark lower in treatment 1 (blind re-mark by assistant examiners) and three marks higher in treatment 2 (blind re-mark by senior examiners). With regard to treatment 3 (non-blind re-mark), the mean is quite close to the original, being only half a mark higher. The standard deviation of the re-marks is smaller than the one in the original marks. The minimum and the maximum marks are similar in all marking treatments.

Table 9 : Summary statistics of the marks awarded in the different marking treatments

	N	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
Original	200	22.08	7.78	4	40
Treatment 1	200	21.53	6.89	5	38
Treatment 2	200	25.31	7.31	6	40
Treatment 3	200	22.73	7.62	4	39

Table 10 displays the absolute differences between the original marks and the three treatments. The average mark change between the original scripts and treatment 1 is 4.49. For treatment 2, the mean is 5.64, which is bigger than for the other treatments. The smallest value corresponds to the non-blind marking (third treatment), where the minimum difference, 0, was achieved in 71 cases. This table provides confirmation of the hypothesis that the marking of two examiners would be affected by whether or not previous marks and comments had been removed from the scripts. Annotations might affect to what exactly within an answer a subsequent examiner will pay attention. Something marked up by the first examiner might be emphasised to a second examiner when they might not have noticed it themselves and, if the first examiner missed something salient, the second examiner may be more likely to do so too (Wilmut, 1984).

Table 10 : Absolute differences in marks

	Mean Difference	*Standard Deviation*
Original – Treatment 1	4.49	3.68
Original – Treatment 2	5.64	4.19
Original – Treatment 3	1.84	2.20

The percentages of exact agreement between the original marks and the different sets of re-marks are 8%, 3% and 36%, respectively. Figures in Table 11 provide evidence of much wider disagreement (in total marks) between English examinations than between Classical Greek examinations. This is no doubt related to the nature of the English examination questions, which are much less tightly structured, allowing for greater freedom in composing a response and requiring more subjective judgement by markers.

Table 11 : Distribution of differences between original and experimental marks

Difference in marks	*Treatment 1 (%)*	*Treatment 2 (%)*	*Treatment 3 (%)*	*Total (%)*
< -13	1.0	4.5	0.0	1.8
-13 to -11	1.5	5.5	0.0	2.3
-10 to -8	6.5	15.0	2.5	8.0
-7 to -5	8.0	15.0	5.0	9.3
-4 to -2	18.5	23.0	20.5	20.7
-1	7.5	5.5	15.5	9.5
0	8.0	2.5	35.5	15.3
1	8.5	6.5	7.5	7.5
2 to 4	19.5	11.1	9.5	13.4
5 to 7	10.0	6.5	3.0	6.5
8 to 10	6.0	4.0	1.0	3.7
11 to 13	3.0	1.0	0.0	1.3
> 13	2.0	0.0	0.0	0.7

Figure 1 illustrates the marks awarded in the three different treatments and the original marks and Pearson's correlation coefficients are displayed in Table 12.

Figure 1 permits a comparison to be made between the marks awarded to the scripts in the different treatments. It can be seen that the variations between the markers' judgements were considerably reduced when they were marking scripts with the original marks and comments on them.

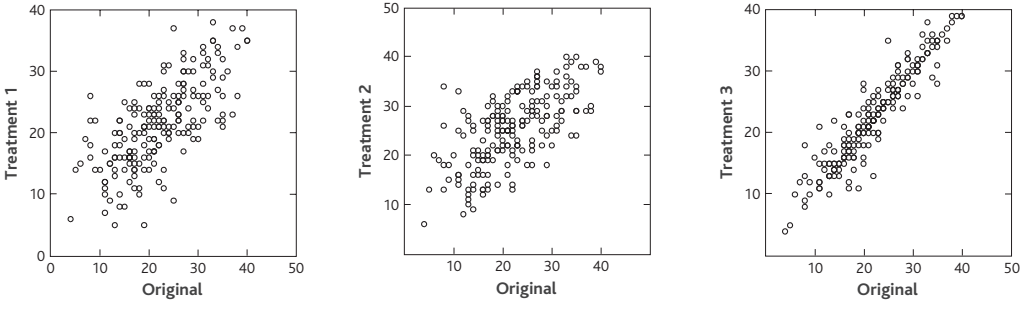

Figure 1 : Scatter diagrams illustrating the relationship between the marks awarded in the different treatments

The correlation coefficients with the original marks are not very high for treatments 1 and 2 indicating that, to a certain extent, the re-markers do not agree closely with the original marks. They also do not agree with one another. The highest correlation appears between the original marks and treatment 3. The correlation between treatment 2 and treatments 1 and 3 is higher than the correlation between treatment 2 and the original marks.

Table 12 : Pearson's correlation coefficients

	Original	*Treatment 1*	*Treatment 2*	*Treatment 3*
Original	1.0000	0.6951	0.6593	0.9346
Treatment 1	0.6951	1.0000	0.6789	0.7417
Treatment 2	0.6593	0.6789	1.0000	0.7276
Treatment 3	0.9346	0.7417	0.7276	1.0000

In terms of the Kappa statistic, for the first treatment we obtain a moderate agreement with the original marks (0.4908). For the second treatment, the value of Kappa, 0.4371, indicates moderate to poor agreement. The level of agreement is higher for treatment 3, with a value of Kappa of 0.7783 (similar to the blind re-mark in Classical Greek), which is a sign of a good agreement.

Reconciliation

Scripts needing reconciliation were determined using the 10% criterion. In this case, reconciliation is performed if the difference in marks is bigger than 4. Table 13 displays the numbers and percentage (in brackets) of scripts that needed reconciliation.

Table 13 : Scripts that needed reconciliation

		Examiner Experience		*Marking*	
	Total	*Blind Assistant*	*Blind Senior/ Experienced*	*Blind*	*Non-blind*
Reconciliation	202 (33.7)	76 (38.0)	103 (51.5)	179 (44.8)	23 (11.5)

In English, the number of scripts needing reconciliation was much higher than for Classical Greek. 202 of the re-marked scripts needed reconciliation. Among them, 76 scripts were blind re-marked by assistant examiners and 103 by senior examiners. 23 scripts that were non-blind re-marked needed reconciliation.

In the three treatments, reconciliation generally provides different outcomes than averaging two marks (see Table 14) and increases the correlation with the original marks and the blind re-marking. Cresswell (1983) demonstrated that the simple addition of the two markers' scores will rarely produce a composite score with the highest reliability possible.

If we had reduced the cut-off point for reconciliation to +/- 2 marks (5% of the mark range) then the reconciliation task would have become enormous. The total percentage of scripts needing reconciliation would have been around 50%. 64% of the blind re-marked scripts and 22% of the non-blind re-marked scripts would have had to be reconciled, greatly increasing costs.

Table 14 : Difference between the mean of two marks and the reconciliation outcome

	Treatment 1		*Treatment 2*		*Treatment 3*	
Difference	*Frequency*	*%*	*Frequency*	*%*	*Frequency*	*%*
-6	1	1.3	2	1.9	1	4.3
-5	1	1.3	5	4.8	3	13.0
-4	2	2.6	4	3.9	3	13.0
-3	4	5.3	12	11.6	1	4.3
-2	13	17.1	16	15.5	0	0.0
-1	10	13.2	10	9.7	0	0.0
0	11	14.8	12	11.6	0	0.0
1	12	15.8	5	4.8	4	17.4
2	8	10.5	14	13.6	5	21.7
3	10	13.2	15	14.6	4	17.4
4	1	1.3	3	2.9	1	4.3
5	2	2.6	4	3.9	1	4.3
6	0	0.0	1	1.0	0	0.0
7	1	1.3	0	0.0	0	0.0

Conclusions and discussion

A first conclusion that can be drawn from this study is that there is a contrast between Classical Greek and English, the former being more reliably marked. Newton (1996) found the same type of contrast between Mathematics, traditionally the most reliably marked subject, and English.

Although in Classical Greek some of the questions required relatively extended answers, the task of the examiners was to award a mark for a specified response. In English, the examiners' task was generally to evaluate the quality of the work. This involved more interpretation and therefore more scope for a difference in opinion.

The results of this investigation appear to provide evidence that removing previous marks and comments from scripts does make a difference. It would seem that examiners who are asked to re-mark scripts cannot help but be affected by previous judgements: the non-blind re-markers can see why the original mark was awarded and they might be happy to concur. Had the second examiners marked blind, they might well not have spotted features noted by the original examiners but also might have spotted features not noted by the original examiners. However, had they marked non-blind, they might have been influenced by incorrect marks or annotations.

There is a need for further research into non-blind double marking. It is necessary to be sure that the second marker will always have the confidence to reject the influence of the marking or the annotations.

One serious impediment to double marking is the increase in administrative time and costs which it entails. Feasibility is a further issue due to the shortage of specialist markers in the UK.

Finally, it should be pointed out that the marking carried out in this research is done under experimental conditions. In the live marking of the examinations, a standardisation meeting is held in order to establish a common standard that is used to maintain the quality of marking during the marking period. Although in this research a meeting with the examiners took place before the re-marking and the principal examiners reviewed the mark schemes with the examiners in order to identify any marking issues, there was no full standardisation meeting. Also, in the live marking period, when the examiners are submitting their marks there are a series of quality control procedures, for example, monitoring the marking, adjustments to the marks of an individual examiner or clerical

checks (details on the quality assurance procedures can be found in QCA Code of Practice, 2005). In this research we examined the marks without performing these procedures.

References

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies*, **52**, 29-46.

Cannings R., Hawthorne K., Hood K. & Houston H. (2005). Putting double marking to the test: a framework to assess if it is worth the trouble. *Medical Education*, **39**, 299-308.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 7-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213- 220.

Cresswell M.J. (1983). *Optimum weighting for double marking procedures*. AEB Research Report, RAC281.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

Lucas, A.M. (1971). Multiple marking of a matriculation Biology essay question. *British Journal of Educational Psychology*, **41**, 78-84.

Massey, A. & Foulkes, J. (1994). Audit of the 1993 KS3 Science national test pilot and the concept of quasi-reconciliation. *Evaluation and Research in Education*, **8**, 119-132.

Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, **22**, 405-420.

Partington, J. (1994). Double marking students' work. *Assessment and Evaluation in Higher Education*, **19**, 57-60.

Pilliner, A.E.G. (1969). Multiple marking: Wiseman or Cox? *British Journal of Educational Psychology*, **39**, 313-315.

QCA (2005). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice 2005/2006*. London: Qualifications and Curriculum Authority.

Uebersax, J. (2003). *Statistical Methods for the Analysis of Agreement*. http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm. Accessed April 2006.

White, R. (2001). Double marking versus monitoring examinations. *Philosophical and Religious Studies*, **1**, 52-60.

Wilmut, J. (1984). *A pilot study of the effects of complete or partial removal of marks and comments from scripts before re-marking them*. AEB Research Report RAC315.

Wood, R. & Quinn, B. (1976). Double impression marking of English language essays and summary questions. *Educational Review*, **28**, 229-246.

Item-level examiner agreement

Nicholas Raikes and Alf Massey Research Division

Abstract

Studies of inter-examiner reliability in GCSE and A-level examinations have been reported in the literature, but typically these focused on paper totals, rather than item marks. See, for example, Newton (1996). Advances in technology, however, mean that increasingly candidates' scripts are being split by item for marking, and the item-level marks are routinely collected. In these circumstances there is increased interest in investigating the extent to which different examiners agree at item level, and the extent to which this varies according to the nature of the item.

Here we report and comment on intraclass correlations between examiners marking sample items taken from GCE A-level and IGCSE examinations in a range of subjects.

The article is based on a paper presented at the 2006 Annual Conference of the British Educational Research Association (Massey and Raikes, 2006).

Introduction

One important contribution to the reliability of examination marks is the extent to which different examiners' marks agree when the examiners mark the same material. Without high levels of inter-examiner agreement, validity is compromised, since the same mark from different examiners cannot be assumed to mean the same thing. Although high reliability is not a sufficient condition for validity, the reliability of a set of marks limits their validity.

Research studies have in the past investigated inter-examiner reliability, but typically these focussed on agreement at the level of the total mark given to scripts. The operational procedures followed by examination Boards for documenting examiner performance also often involve recording details of discrepancies between examiners at the script total level. New technologies are facilitating new ways of working with examination scripts, however. Paper scripts can now be scanned and the images transmitted via a secure Internet link to examiners working on a computer at home. Such innovations are creating an explosion in the amount of item-level marks available for analysis, and this is fostering an interest in the degree of inter-examiner agreement that should be expected at item level. The present article provides data that will help inform discussions of this issue.