Item-Level Examiner Agreement

A. J. Massey and Nicholas Raikes*

Cambridge Assessment, 1 Hills Road, Cambridge CB1 2EU, United Kingdom

*Corresponding author

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.

In the present paper 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. We also consider whether any simple relationship exists between the size of intraclass correlations and surface features of items such as: the type and length of response sought; the implied time restriction imposed on candidates; the range and organisation of marking points within the item; the nature of the mark scheme and the extent markers are permitted to exercise discretion.

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 of script totals. 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. We have published the present paper to provide data that will help inform discussions of this issue.

Although a rigorous approach to predicting expected levels of inter-examiner agreement would consider the specific cognitive demands placed on examiners by marking particular items under particular circumstances, such an approach would be impractical in operational settings. We have therefore considered surface features of the items and their mark schemes that might be expected to influence the reliability with which they are marked. We have included data concerning these contextual features with our results, so that any patterns may become apparent. The surface features considered are:

(i) The subject being examined;

(ii) The level of the examination;

(iii) The maximum mark available for the item;

(iv) The implied time restriction (ITR) imposed on candidates. This is:

$$\text{Total time in minutes} \quad X \quad \frac{\text{Item maximum mark}}{\text{Total maximum mark}}$$

(v) The type of marking employed, namely "objective", "points based" or "levels based", as follows:

Objective marking – items that are objectively marked require very brief responses and greatly constrain how candidates must respond. Examples include items requiring candidates to make a selection (e.g. multiple choice items), or to sequence given information, or to match given information according to some given criteria, or to locate or identify a piece of information (e.g. by marking a feature on a given diagram), or to write a single word or give a single numerical answer. The hallmark of objective items is that all credit-worthy responses can be sufficiently pre-determined to form a mark scheme that removes all but the most superficial of judgements from the marker.

Points based marking – these items generally require brief responses ranging in length from a few words to one or two paragraphs, or a diagram or graph, etc. The key feature is that the salient points of all or most credit-worthy responses may be pre-determined to form a largely prescriptive mark scheme, but one that leaves markers to locate the relevant elements and identify all variations that deserve credit. There is generally a one-to-one correspondence between salient points and marks.

Levels based marking – often these items require longer answers, ranging from one or two paragraphs to multi-page essays or other extended responses. The mark scheme describes a number of levels of response, each of which is associated with a band of one or more marks. Examiners apply a principle of best fit when deciding the mark for a response.

(vi) For levels based marking, the number of levels available.

These factors are related to each other. For example, the Implied Time Restriction is related to the kind of marking employed, since objective marking is generally used for the quickest items and levels-based marking for items that require candidates to generate long responses.

The source of our data

The analysis presented in the present paper was of data collected during trials of new ways for examiners to record item-level marks. All marking for the trials was done using paper scripts (i.e. no marking was done on screen).

Data for one component from each of five subjects were available as follows:

- IGCSE Foreign Language French: Listening – multiple choice and short, textual-answer items worth one or two marks. Total mark = 48. Time = 45 mins;
- IGCSE Development Studies: Alternative to Coursework – short answer items worth between one and six marks. Total mark = 35. Time = 90 mins;
- A-Level Chemistry: Structured Questions – multiple choice and short answer items worth between one and five marks. Total mark = 60. Time = 60 mins;
- A-Level Economics: Data Response and Case Study – short, textual answer items worth between one and six marks, plus some longer textual answer items worth between eights and twelve marks. Total mark = 50. Time = 110 mins;

- A-Level Sociology: Principles and Methods – candidates chose two from six twenty-five mark essay items. Total mark = 50. Time = 90 mins.

For each Component, a sample of 300 scripts was drawn, covering a range of ability and countries. All scripts came from centres with twenty to forty candidates. All marks and examiner-annotations were removed and the scripts were copied.

Three examiners from each component were recruited for the study. One examiner each from Sociology and Chemistry dropped out. Each remaining examiner marked the 300 copied scripts after live, operational marking had been completed. The examiners recorded their item marks using the methods that were being trialled. The live item marks were also keyed from the original scripts. We therefore had up to four independent sets of marks for the 300 scripts from each component.

Caveats: the study examiners were marking non-live, and inter-examiner agreement levels may be different in live situations. The study examiners were also recording their marks using two different methods, which may have had a small impact on their reliability.

Results

Pearson correlations for script totals

Although item-level data are the main focus of the present paper, we present Pearson correlations for script totals in Tables 1 to 5. The correlations are in line with those found for similar subjects in other studies. For example, Newton (1996) reported mark- re-mark correlations of between 0.992 and 0.997 for GCSE mathematics, and 0.87 and 0.95 for GCSE English. Murphy (1976), quoted in Newton (1996), found mark- re-mark correlations of 0.73, 0.76 and 0.85 for three A-Level English components. The reliability of the marking reported in the present paper does not therefore appear to be atypical.

Table 1: Pearson correlations for script totals – IGCSE French Listening

	Exr F1	**Exe F2**	**Exr F3**	**Exr F4**
Exr F1	1.000	0.995	0.994	0.995
Exr F2	0.995	1.000	0.994	0.996
Exr F3	0.994	0.994	1.000	0.994
Exr F4	0.995	0.996	0.994	1.000

Table 2: Pearson correlations for script totals – IGCSE Development Studies

	Exr DS1	**Exr DS2**	**Exr DS3**	**Exr DS4**
Exr DS1	1.000	0.905	0.939	0.899
Exr DS2	0.905	1.000	0.920	0.923
Exr DS3	0.939	0.920	1.000	0.934
Exr DS4	0.899	0.923	0.934	1.000

Table 3: Pearson correlations for script totals – A-Level Chemistry

	Exr C1	Exr C2	Exr c3
Exr C1	1.000	0.993	0.991
Exr C2	0.993	1.000	0.991
Exr C3	0.991	0.991	1.000

Table 4: Pearson correlations for script totals – A-Level Economics

	Exr E1	Exr E2	Exr E3	Exr E4
Exr E1	1.000	0.822	0.804	0.696
Exr E2	0.822	1.000	0.744	0.664
Exr E3	0.804	0.744	1.000	0.743
Exr E4	0.696	0.664	0.743	1.000

Table 5: Pearson correlations for script totals – A-Level Sociology

	Exr S1	Exr S2	Exr S3
Exr S1	1.000	0.830	0.920
Exr S2	0.830	1.000	0.850
Exr S3	0.920	0.850	1.000

Item-level Intraclass Correlations (ICCs)

In this section we report the intraclass correlation coefficients for each item within each component. The intraclass correlation may be interpreted as the proportion of variance in the set of candidates' marks that is due to the candidates, i.e. after examiner effects have been controlled for. That is, if there is perfect agreement between the examiners on every script, the intraclass correlation coefficient will be 1; but if there is no agreement and the marks appear random, the coefficient will be 0. We have chosen to report intraclass correlations, rather than Pearson correlations, because the intraclass correlation reflects the degree of agreement between two or more examiners, whereas the Pearson correlation reflects the extent to which the relationship between two examiners' marks is linear – a high Pearson correlation would be obtained even if one examiner was consistently more or less severe than the other. The intraclass correlations were calculated by SPSS version 12 for Windows, using a two-way random consistency model, and the values reported are the single measures. The tables contain the following information:

Item	A label for the item
Max	Maximum mark available for the item
ITR	Implied Time Restriction in minutes

The remaining columns give single-measure, consistency intraclass correlation coefficients for:

Obj	Objective items
Px	Items with a points-based marking scheme, where x denotes the maximum

number of points to be credited in any response.

Ln-m	Items with a levels marking scheme, where n denotes the number of levels and m denotes the total number of marks available

The bottom rows of each table give:

Mean raw r	The mean intraclass correlation for the column
Mean adj r	Raw correlations are not strictly additive and their mean may be biased towards zero. Because of this the individual intraclass correlations were converted to additive values using Fisher's r to z transformation[1]. The mean of these values for each column was calculated and converted back into an r, presented in this row. NB we have included this value for completeness, but in the current context we prefer the mean of the raw values since the r to Z transformation gives great weight to outliers approaching 1. Indeed on two instances we have correlations of 1 and the corresponding z value is therefore incalculable; very different values are obtained if, for example, 0.99 or 0.999 or 0.9999 is substituted.
n_{items}	The number of items in the column

Below each table we give

$n_{scripts}$	The number of scripts marked by each examiner (approximately 300, but some scripts were excluded because one or more of the trial examiners neglected to mark it – of course all scripts were live marked).
$n_{examiners}$	The number of examiners who marked each response
ITR per mark	The Implied Time Restriction per mark
r_{tot}	The mean intraclass correlation between the examiners' total marks for the scripts.

[1] Fisher's r to z transformation: $z' = .5[\ln(1+r) - \ln(1-r)]$

Table 6: Item-level ICCs – IGCSE French Listening

Item	max	ITR	Obj	P1	P2
1	1	0.9	0.972		
2	1	0.9	0.975		
3	1	0.9	0.982		
4	1	0.9	0.992		
5	1	0.9	0.995		
6	1	0.9	0.989		
7	1	0.9	0.978		
8	1	0.9	0.972		
9	1	0.9	0.989		
10	1	0.9	0.974		
11	1	0.9	0.956		
12	1	0.9		0.934	
13	1	0.9	0.867		
14	1	0.9	0.963		
15	1	0.9	0.977		
16	1	0.9	0.991		
17	6	5.6	0.986		
18	1	0.9		0.928	
19a	1	0.9		0.961	
19b	1	0.9		0.809	
20	1	0.9		0.963	
21	1	0.9		0.881	
22	1	0.9		0.912	
23	1	0.9		0.675	
24	1	0.9		0.894	
25	2	1.9			0.821
26	1	0.9	0.984		
27	1	0.9	0.992		
28	1	0.9	0.999		
29	1	0.9	0.964		
30	1	0.9	0.994		
31	1	0.9	0.969		
32	2	1.9			0.919
33	1	0.9		0.931	
34	1	0.9		0.906	
35	1	0.9		0.799	
36	2	1.9			0.815
37	1	0.9			0.881
38	1	0.9			0.753
39	1	0.9			0.926
		Mean raw r	0.975	0.883	0.853
		Mean adj r	0.984	0.903	0.865
		n_{item}	22	12	6

$n_{scripts}$: 300 $n_{examiners}$: 4 ITR per mark: 0.9 r_{tot}: 0.995

The French Listening paper contained more objectively marked items and the shortest Implied Time Restriction per mark of any of the papers considered. The ICC for the script totals was very high (0.995), indicating a very high degree of agreement overall. The average item ICC was highest for the objective items, slightly lower for the one-mark points-based items, and a little lower still for the two-mark points-based items.

Table 7: Item-level ICCs – IGCSE Development Studies

item	max	ITR	Obj	P1	P2	P3	P6	L4-4
1ai	1	2.6	0.981					
1aii	2	5.1			0.401			
1aiii	2	5.1	0.978					
1aiv	2	5.1			0.661			
1av	1	2.6		0.852				
1bi	2	5.1			0.602			
1bii	2	5.1			0.773			
1ci	1	2.6		0.914				
1cii	3	7.7				0.624		
1ciii	3	7.7				0.824		
1di	4	10.3						0.890
1dii	3	7.7				0.716		
2a	3	7.7				0.712		
2b	6	15.4					0.809	
Mean raw r			0.980	0.883	0.609	0.719	0.809	0.890
Mean adj r			0.980	0.887	0.627	0.727	0.809	0.890
n_{item}			2	2	4	4	1	1

$n_{scripts}$: 265 $n_{examiners}$: 4 ITR per mark: 2.6 r_{tot}: 0.917

Development Studies had the longest ITR per mark of any of the papers, a reflection, perhaps, of the amount of writing that candidates were expected to do and the fact that IGCSE is an international examination aimed at 14 to 16 year olds. The ICC for the script totals (0.917) was high. The mean ICCs for the objective and one-mark points-based items were very similar to those for French Listening, but it was considerably lower for the two-mark items. The mean ICC was higher for the P3 items than for the P2 items, and the ICC was also quite high for the P6 item and the four mark levels-based item. There is therefore not a simple relationship between the number of marks available for an item and the intraclass correlation.

Table 8: Item-level ICCs – A-Level Chemistry

item	max	ITR	Obj	P1	P2	P3
1a	1	1		0.737		
1b	1	1		0.894		
1ci	1	1	0.992			
1cii	1	1	0.993			
1d	2	2			0.927	
1e	3	3				0.939
1fi	1	1		0.800		
1fii	1	1		0.840		
1fiii	1	1		0.857		
2ai	1	1		0.918		
2aii	1	1		0.850		
2aiii	1	1		0.924		
2aiv	1	1		0.727		
2b	1	1		0.949		
2ci	1	1		0.894		
2cii	2	2			0.828	
2ciii	1	1		0.861		
2civ	1	1		0.799		
3ai	2	2			0.917	
3aii	2	2			0.837	
3bi	1	1		0.757		
3bii	1	1		0.638		
3biii	3	3				0.844
3ci	2	2			0.775	
3cii	1	1		0.663		
4a	2	2		0.963		
4bi	1	1		0.896		
4bii	2	2			0.963	
4biii	1	1		0.935		
4ci	1	1		0.866		
4cii	2	2			0.823	
5a	2	2			0.957	
5b	2	2			0.884	
5c	2	2			0.856	
5d	1	1		0.901		
5ei	1	1	0.866			
5eii	1	1		0.851		
5fi	1	1		0.966		
5fii	2	2			0.884	
6ai	1	1		0.824		
6aii	1	1		0.891		
6b	1	1		0.696		
6c	3	3			0.913	
		Mean raw r	0.950	0.842	0.880	0.892
		Mean adj r	0.980	0.866	0.894	0.902
		n_{item}	3	26	12	2

$n_{scripts}$: 298 $n_{examiners}$: 3 ITR per mark: 1 r_{tot}: 0.992

The Chemistry paper had the second shortest ITR per mark and the second highest ICC for the script totals. The mean ICC for the two objective items (0.950) was

higher than for any of the other categories. The mean ICC dropped to 0.842 for the P1 items, but then rose for the P2 and P3 items.

Table 9: Item-level ICCs – A-Level Economics

item	max	ITR	Obj	P2	P3	P6	L3-8	L3-10	L3-12
1ai	1	2.1	0.978						
1aii	2	4.2		0.879					
1bi	3	6.3			0.489				
1bii	2	4.2		0.668					
1ci	3	6.3			0.554				
1cii	3	6.3			0.507				
1d	6	12.6				0.548			
2a	8	16.8					0.740		
2b	10	21.0						0.567	
2c	12	25.2							0.585
Mean raw r			0.978	0.774	0.517	0.548	0.740	0.567	0.585
Mean adj r			0.978	0.797	0.517	0.548	0.740	0.567	0.585
n_{item}			1	2	3	1	1	1	1

$n_{scripts}$: 294 $n_{examiners}$: 4 ITR per mark: 2.1 r_{tot}: 0.774

The A-Level Economics paper had the lowest script-total ICC (though at 0.774 this is still respectable) and the second longest ITR per mark. The ICC for the objective item was very high (0.978); the mean ICC for the two P2 items was substantially lower at 0.774; and lower still for the P3 items (0.517). The ICCs for the P6 item and the 10- and 12-mark 3-Levels items were a little higher, but similar. The 8-mark 3-level item appears anomalous with its ICC of 0.740, and no simple explanation is obvious – the item itself appears straightforward enough, asking candidates to explain a possible link between interest rates and inflation.

Table 10: Item-level ICCs – A-Level Sociology

item	max	ITR	L4-25
a1	25	45	0.865
a2	25	45	0.851
b3	25	45	0.874
b4	25	45	0.795
c5	25	45	0.767
c6	25	45	0.797
Mean raw r			0.825
Mean adj r			0.829
n_{item}			6

$n_{scripts}$: 252 $n_{examiners}$: 3 ITR per mark: 1.8 r_{tot}: 0.863

Candidates had to write two extended essays for this A-Level Sociology paper, with an ITR per essay of 45 minutes – though the ITR per mark is actually lower than for Economics or Development Studies. The ICCs for the items are quite high and very similar, ranging from 0.767 to 0.865, with a mean of 0.825.

Discussion

There is quite a strong relationship between the Implied Time Restriction per mark and the total mark intraclass correlations – the Pearson correlation between these quantities is -0.66, rising to a very high -0.99 if Development Studies, with its apparently generous time restriction, is excluded. This is probably because there is a positive relationship between the amount candidates are expected to write per mark and the amount of time they are given to write it, and there is more scope for examiners to disagree on longer answers.

Turning to the item-level results, the objective items were marked very reliably regardless of the subject – the mean ICC was 0.95 or higher for all four of the subjects that had objective items.

The mean ICC for the P1 items, where candidates typically had to write a few words, was 0.883 for both French Listening and Development Studies, and 0.842 for Chemistry.

For P2 items, the mean ICC ranged from 0.609 (Development Studies) to 0.880 (Chemistry). However, there is a strong relationship between the Implied Time Restriction and the mean ICC (the Pearson correlation is -0.93).

The mean ICC for the P3 items was even more variable, ranging from 0.517 (Economics) to 0.892 (Chemistry). There were two P6 items: Development Studies (ICC = 0.809) and Economics (ICC = 0.548).

There is no simple relationship between the mean ICC and the maximum number of marks available for points based items. In French Listening, the mean ICC is a little lower for P2 items than for P1 items; conversely, in Chemistry it is a little higher for P2 items than for P1 items, and a little higher still for P3 items. In Development Studies and Economics, the mean ICC drops then rises with increasing maximum marks.

Turning to the Levels-based items, the four-level, four mark Development Studies item had an ICC of 0.890, considerably higher than the ICCs for any of this paper's P2 or P3 items, or for the P6 item. The mean ICC for the four-level, 25-mark Sociology items was high, at 0.863. The ICC for the three-level, eight mark Economics item was 0.740, but the other two three-level Economics items had considerably lower ICCs – although both were higher than the mean ICC for the P3 Economics items. Given the relatively low ICCs in Economics generally, these two items may be anomalous, and the ICC for the eight mark item may be more typical of levels-based items generally. The Economics levels-based items were less open-ended than the Sociology items, and the marking scheme contained far more content. Sanderson (2001), in his work about A-Level essay marking, concludes that there is no guarantee that highly specified mark schemes are used to the full, and 'given the limits of working memory, complex marking schemes may inhibit the development of an accurate representation of candidate texts on the part of examiners' (p. 279). This might help explain why the Economics marking was so much more variable than the Sociology marking, though this suggestion is speculative.

To conclude, in this paper we have presented detailed information about inter-examiner agreement at both whole script and item level in IGCSE and A-Level examinations in a range of subjects. We found a strong negative relationship between the Implied Time Restriction per mark and total-mark intraclass correlations. We considered whether there was a relationship between surface features of the items, particularly the type of mark scheme used, and the mean intraclass correlation for the items. The results were mixed. Mean ICCs for objective items were 0.95 or

higher for all four subjects that had them, and were well above 0.80 for P1 items for the three subjects that had P1 items (these are one-mark items where candidates typically have to write a few words). Mean ICCs were more variable for higher tariff items marked using a points-based marking scheme. Items marked against levels-based marking schemes generally had quite high ICCs, in most cases ranging from 0.740 to 0.890, though two Economics items had much lower ICCs. Relatively few levels based items were considered, however. We recommend that more data be collected and analysed.

References

Murphy, R.J.L. (1978). 'Reliability of marking in eight GCE examinations', *British Journal of Educational Psychology*, **48**, 2, 196–200.

Newton, P.E. (1996). 'The Reliability of Marking of General Certificate of Secondary Education Scripts: mathematics and English', *British Educational Research Journal*, **22**, 4, 405-420

Sanderson, P. J. (2001). 'Language and Differentiation in Examining at A level', *PhD thesis*, University of Leeds.