

**A CONCURRENT APPROACH TO ESTIMATING THE RELIABILITY
OF ELECTRONIC MARKING OF LONG FORM ANSWERS**

ABSTRACT

The reliability of electronic marking was investigated for a sample of GCSE History scripts in a trial of DRS's current long form answer version of the CMI+ software. The concurrent approach adopted for the investigation met procedural difficulties, and the planned quantity of research data could not be collected. The segmentation that characterises e-marking was fully exploited in the data analyses, but they could be completed for only a small number of examiners and scripts. While the analyses suggested both low marking reliability, based on absolute mark differences from the Principal Examiner's (concurrent) 'true' marks, and high marking reliability, based on correlation coefficients, the evidence is far too slight to determine whether or not marking reliability is at expected levels. A recommendation of this project is therefore that more substantial evidence from a variety of papers should be sought from future series, with further research to explore the reliability of screen-based marking of long form answers as more examiners and more components are brought into the process. The 'S portion' approach to data collection is commended: it is a valuable method, unique to electronic marking, for providing good quality concurrent data for investigating the reliability of on-screen marking and checking on the suitability of the variety of long form answer components that AQA hopes to draw into electronic marking.

1. INTRODUCTION

The newly developed version of DRS's electronic marking software CMI+ to cater for 'long form answers' (LFAs) was trialled with three components in the June 2009 series. The components were two GCE Classical Civilisation AS units (CIV1A-F and CIV2A-F, from the new specification) and a single GCSE History component (Specification B, Paper 3). Each offers candidates a substantial choice of options within the papers, plus some further choice within the selected options. Candidates used a generic answer booklet in which they identified their chosen questions by copying their three-digit question numbers into pre-printed boxes in the left hand margins. This is the mechanism that has been developed so that the marking software can accommodate question choice and variable length written responses. Recognition of the scanned question numbers and the resulting allocation of responses to questions by the software obviously depend on accurate and clear identification of questions by candidates.

A further key feature of the LFA software is that the standard CMI+ practice of monitoring marking with seeds can be replaced by a peer double-marking monitoring regime, intended for questions that have a maximum mark of 12 or more. Also available as an additional feature is a standardisation facility, in which the Principal Examiner provides marks for a number of responses which are referred to 'S portions'. For purposes of standardisation the S portions would be marked by examiners at the beginning of their marking of each question. Although not being used by AQA for this purpose, this standardisation facility can be made available by DRS

for research or other use, and it permits the design of a concurrent rather than the more usual post-results approach to investigating e-marking reliability (a desirable feature in planning reliability studies in a fast moving area of development). This report gives details of an investigation based on S portions that was planned to permit comparisons of examiners' live marking with the marking of the Principal Examiner and the senior team, using a selection of S portions delivered at regular intervals through the examiners' marking. The investigation involved just the GCSE History component.

In previous marking reliability studies, for example investigations of marking in GCSE English (Fowles, 2006a, 2006b and 2009), and as discussed by Meadows and Billington (2005), the Principal Examiner's marks are defined as the 'true' marks, and comparisons are made of examiners' marks with the true mark for each complete paper. Variability in total marks for the same scripts provides an indication of marking reliability: perfectly reliable marking requires that the same marks are given on any marking occasion, while differences are the basis for estimating the degree of marking reliability.

One thread of e-marking reliability research is designed to examine the transition from paper-based to electronic marking, to compare the outcomes from the different marking media and procedures. For this purpose the same script samples must be marked in both media. However the current investigation was not designed with a view to providing a direct link to paper-based marking. It is instead capitalising on the facility in electronic marking to provide concurrent marking data for the same scripts from every member of a team of examiners, without their awareness of which scripts are being used to generate the data – a luxury in paper-based marking that is operationally unobtainable.

2. METHOD AND DATA COLLECTION

2.1 Design

The plan for this investigation of marking reliability made use of the standardisation feature in DRS's 'CMI+ L' software in which particular responses are identified as 'S portions' and marked by every examiner before they embark on their live marking. It was agreed with DRS that the choice of S portions for the purposes of investigating marking reliability (rather than for standardisation of marking) would be the complete set of responses in a small sample of the GCSE History Paper 3 scripts. The S portions would be made available for marking by the examiners at various points in the marking period, having been marked at the outset (pre-standardisation) by the Principal Examiner. Each S portion would then be marked by each examiner on the examining panel, and the marks of each examiner aggregated into a set of total marks for each sampled script. By this means the variability of script totals could be examined and the degree of marking reliability assessed.

The S portions would only be used for investigating marking reliability and would be quite separate from the 10 per cent or so of double marked 'benchmark portions' included for monitoring purposes in an examiner's allocation. Being designated as 'research' portions, the S portions would make no contribution to that monitoring, nor serve any standardising function.

2.2 Script selection for use as S portions

The format of the GCSE History paper used in the investigation differed from its format in previous series. It used a new style question numbering system that needs a brief introduction. The traditional numbering system presents difficulties for accurate character recognition so, in the LFA numbering system designed for this first trial, each question part was identified by three digits (the system has since been replaced by a 2-digit system). Thus a Section A part question

previously referred to as Question 1(a) was presented on the 2009 question paper as question '111'.

The question paper gives candidates a choice of one question from three in each of its two Sections A and B, each question addressing a different optional topic from the specification. Each Section A question is made up of two compulsory parts and a third part chosen from two, while each Section B question is made up of three compulsory parts and a fourth part chosen from two. Each question has a total of 30 marks, giving the whole paper a total mark of 60. In preparing the paper for e-marking, it was agreed at the suggestion of the Principal Examiner that the first two parts of each Section A question should be linked for marking, that is, although they would be marked as separate items, they would be clipped together in scanning so that the same examiner would mark the pair of responses. The same arrangement was also agreed for the first three parts of each of the Section B questions. The question paper structure, e-marking numbering scheme, clip linkages and option details are summarised below.

Section	Qn	'Old' qn part	'New' qn part	Mark allocation		Option topic (and % choosing ¹)
A	1	(a)	111	5	linked	1 The changing role and status of women in Britain since 1900 (76.9%)
		(b)	112	10		
		(c) (i)	113	15		
		(ii)	114	15		
	2	(a)	121	5	linked	2 Britain and Ireland since 1916 (12.6%)
		(b)	122	10		
		(c)(i)	123	15		
		(ii)	124	15		
	3	(a)	131	5	linked	3 Britain's changing role in the world since 1956 (10.5%)
		(b)	132	10		
		(c)(i)	133	15		
		(ii)	134	15		
B	4	(a)	241	8	linked	4 Vietnam since 1939 (82.2%)
		(b)	242	6		
		(c)	243	8		
		(d)(i)	244	8		
		(ii)	245	8		
	5	(a)	251	6	linked	5 The Arab Israeli Conflict (1.3%)
		(b)	252	8		
		(c)	253	8		
		(d)(i)	254	8		
		(ii)	255	8		
	6	(a)	261	8	linked	6 Race relations in the USA post 1945 (16.5%)
		(b)	262	6		
		(c)	263	8		
		(d)(i)	264	8		
		(ii)	265	8		

¹ The source of the option choice information is the item analyses provided ahead of the awards meeting by DRS, based in this case on 86% of the candidature.

Thus Section A 'old' Question 1 parts (a) and (b) became 'new' part questions 111 and 112, and they appeared together on the same clip for marking. In summary, each candidate makes seven responses, three from Section A that are marked in two clips and four from Section B that are also marked in two clips, giving a total of four clips.

A common feature of question papers with LFAs is that they offer candidates a degree of choice and, as noted in the Introduction, the new 'CMI+ L' software is designed to accommodate this. In each section of the History paper there is one question which is hugely more popular than the other two, year on year; in Section A it is Question 1 on '*The changing role and status of women in Britain since 1900*' and in Section B it is Question 4 on '*Vietnam since 1939*'.

An initial selection of potential scripts for the reliability study was made by DRS from fast track scripts received and scanned within a couple of days of the examination. These scripts had all been scanned in their entirety, with successful character recognition identifying each part question. A file detailing the electronic location of each script was passed to the Research department for selection of the small study sample. It was decided to confine the sample scripts to those with the particular combination of the two most popular questions (Qn A1 and Qn B4) and also to the more popular of the two question parts within the questions². A systematic selection was made from the DRS file with replacement to meet the question choice restriction and also to limit the number from any particular centre to a maximum of four. The detail of a sample of 33 scripts was passed back to DRS to be made available to the senior examining team online for marking with a view to selecting 30 as suitable for use as S portions.

2.3 Data limitations

The design of the investigation allowed for the Principal Examiner's scrutiny and marking of a sample of at least 30 scripts to take place on the day before examiner standardisation of the eight assistant examiners in the panel, and then for each of the eight to mark all the responses on the sampled scripts, giving a minimum total of 240 script total marks for analysis. He was accompanied by the two Team Leaders to assist in this task as required. However, this was the first time that DRS had used the S portion procedure and there were a few operational complications which had an impact on the output from the senior examiners' initial marking day. As a result not all the scripts that the senior examiners had identified as the sample were found to be marked in their entirety when the parts were subsequently put together, and there were too few spare scripts available to make up for this shortfall. Further, one script was unfortunately later found to have had an amanuensis and had been included twice in the sample (i.e. in both original and amanuensis form). The eventual total of scripts fully marked at this stage and identifiable as the script sample for the investigation, with all responses defined as S portions, was 24.

2.4 Marking

The marking period following standardisation was to some extent problematic. Two of the eight examiners withdrew, and the remainder did not achieve the anticipated rate of marking. Steps were taken to recruit extra examiners and to assist them by offering office-based e-marking, in order to meet the marking schedule. Examiners were allowed to focus on particular items, with the result that coverage of the S portions by the original examiners became patchy. DRS provided a datafile on completion of marking which contained over a thousand item level marks

² The more popular questions were questions 113 and 244. The questions answered in the selected scripts were therefore 111, 112 and 113 (Section A) and 241, 242, 243 and 244 (Section B).

for the S portion responses. Table 1 gives the number of recorded S portion marks by examiner. It shows that the pattern of marking for two of the assistant examiners (Examiners 4 and 5 in the table) were complementary, overlapping fully only on part questions 111 and 112, and full coverage was obtained for only three assistant examiners. It should be noted that the Principal Examiner and Team Leader marks reported in Table 1 are those recorded during concurrent, post-standardisation marking, in response to the concerns over the slow rate of marking, and are not their initial pre-standardisation marks. The Principal Examiner recorded complete concurrent marks for 17 of the 24 scripts. The Team Leaders' concurrent marking was focussed on the first and last part questions so did not give complete script totals. Concurrent marks from only the Principal Examiner and three assistant examiners (Examiners 1 to 3) are therefore included in the analyses that follow.

Table 1 Number of S portion marks by linked part questions and examiner

Part questions	111 - 2	113	241- 3	244
Mark allocations (total 60)	15	15	22	8
Examiners 1, 2 and 3	24	24	24	24
Examiner 4	24	0	24	0
Examiner 5	24	24	4	24
Team Leader 1	24	0	6	24
Team Leader 2	24	0	1	24
Principal Examiner	24	24	17	24

3. ANALYSIS AND RESULTS

3.1 Data

While e-marking is carried out at item level, the focus of reliability estimation is at whole script level since it is variability at candidate level that is at issue in reporting examination results. The S portion marks at item level were therefore aggregated to candidate level, giving script totals which were initially available for analysis in three categories:

- (a) the Principal Examiner (supported by the two Team Leaders) from the pre-standardisation script selection day: all 24 scripts;
- (b) the Principal Examiner, during his live marking, as noted above: 17 scripts;
- (c) three assistant examiners: all 24 scripts.

There are some aspects of the collection of the item marks in category (a) that raise doubts as to the suitability of the resulting script totals as 'true' marks.

- As a newly introduced procedure, DRS met some initial difficulties in delivering the S portions for marking, as already noted. This meant that a shorter and more pressurised time was available for the pre-standardisation marking.
- The senior team had not fully finalised the mark scheme for the paper: the pre-standardisation marking suggested improvements to them as they marked. (The Principal Examiner's concurrent marks in (b) therefore reflect the more up to date interpretation of the mark scheme.)
- The marking at the AQA offices took place in a large room equipped with PCs for general marking, some of which were being used by small groups of examiners in other subjects. This environment was probably more stressful for the panel than had they been able to mark at home to their own timetable.

Table 2 gives details of the script totals in each of the categories (a) to (c), with category (a) included for completeness despite the doubts noted above as to the suitability as 'true' marks of the script totals from the item marks collected in the pre-standardisation marking category. It

will be seen in Table 2 that the marks for scripts 8 and 21 in category (a) show particularly large differences from the other marks in Table 2 (including those of the Principal Examiner for script 8 in (b)). These differences are particularly significant given the grade boundary marks for this paper and the examination in total. The grade boundaries are given in Table 3; the differences between them are all four marks at paper level, and either 16 (grades G to C) or 20 (grades C to A*) marks at subject level.

There is another perspective on the compilation of script totals that suggests using another category of data. Compiling script totals from a single source – whether the Principal Examiner ((a) and (b)) or an assistant examiner ((c)) – is a throw back to the conventional approach to marking scripts on paper, where an individual examiner takes responsibility for marking a whole paper. In electronic marking of course this is not the case; script segmentation means that marking is distributed among several examiners. Segmentation is likely to reduce the impact of the more extreme examiners in terms of their severity or leniency of marking, and there is likely to be a 'regression to the mean' effect.

A further set of script totals has therefore been devised to reflect the advantage given by e-marking over paper-based marking by virtue of segmentation. This further data category involved amalgamating different examiners' marks, drawing on the recorded marks to generate further sets of script totals from the individual examiners' marks. For this purpose the concurrent marks of the Principal Examiner, Examiners 1 to 3 and also Examiners 4 and 5 (see Table 1) were included. Each amalgamated examiner's part question marks drew unsystematically from the available marks, as would be experienced in normal e-marking, but in such a way that no examiner donated more than one question part mark to any script total. This gives a further data category:

(d) 5 'amalgamated' examiners, labelled A1 to A5: 17 scripts

Table 2 includes details of the five script totals in category (d), adding to those obtained in categories (a) to (c). These totals are less varied than those in the earlier categories (each script other than script 16 has its smallest standard deviation for (d)), which reflects a particular benefit of segmented marking and supports the approach of amalgamating or pooling marks from different examiners.

Table 2 Individual script totals by category

script	(a) pre-stand.	(b) PEX, live	(c) 3 examiners		(d) 5 amalgamated examiners	
			mean	sd	mean	sd
1	47	46	45.7	7.6	45.2	4.0
2	57	51	50.7	2.1	51.8	2.3
3	53	53	47.3	3.8	49.2	1.8
4	39	35	37.7	8.1	37.4	4.2
5	33	35	35.7	6.8	34.2	4.8
6	36	42	37.3	6.4	37.6	4.2
7	40	45	44.0	5.0	43.8	2.6
8	13	28	25.3	6.1	24.6	4.4
9	33	36	29.0	5.3	30.2	3.3
10	29	34	34.0	7.2	32.0	3.5
11	22	27	22.7	5.8	23.0	5.2
12	28	36	29.7	4.2	32.2	3.3
13	39	43	37.7	2.9	38.8	2.3
14	30	32	25.3	3.8	27.2	1.9
15	43	52	44.3	5.5	45.4	5.5
16	35	33	32.7	3.1	29.2	6.6
17	44	42	43.0	8.0	41.0	3.7
mean	36.53	39.41	36.59	-	36.64	-
sd	10.86	8.10	8.41	-	8.58	-
18	28	-	27.0	7.9	-	-
19	36	-	39.3	4.5	-	-
20	49	-	48.7	4.0	-	-
21	31	-	19.7	0.6	-	-
22	42	-	38.7	3.5	-	-
23	41	-	35.3	4.2	-	-
24	37	-	34.3	5.7	-	-

Table 3 GCSE History Specification B grade boundary marks

Grade	Paper 3	Subject ³
max	60	200
A*	45	181
A	41	156
B	37	136
C	33	116
D	29	100
E	25	84
F	21	68
G	17	52

³ Note: there is no scaling of Paper 3 marks (maximum 60) in the subject total

3.2 Basis of the script comparisons

Discarding the script totals in the pre-standardisation category (a) restricts comparisons of script totals to 17 scripts, all of which have concurrent Principal Examiner marks (b) which can stand as the 'true' marks. These true marks can be compared with the average of either of two other sets of script total marks, those from:

- the three assistant examiners (c); or
- the five amalgamated examiners' marks (d).

The choice is essentially between the equivalent of paper-based comparisons, based on examiners' whole script totals, and comparisons that exploit the key feature of electronic marking of segmentation and the allocation of question parts to different examiners. The choice that has been taken forward in the comparisons that follow is of the concurrent Principal Examiner marks with the amalgamated examiners' marks⁴. The choice of amalgamated examiners' marks is appropriate because perfectly reliable marking requires that the same total marks are given on any marking occasion – and in the context of segmented e-marking this can be translated as requiring that the same total marks are given by any combination of examiners' part marks.

3.3 Marking reliability

Marking reliability is typically examined from the perspective of the absolute mark difference (AMD) between an examiner's mark and a measure of the script's true mark, and from the correlation between the marks awarded by an examiner and the true mark. The emphasis here is on whole script totals rather than item marks. As already noted, variability at candidate level is at issue in reporting examination results, which is why it is the primary concern of the analysis.

3.3.1 Mark differences in script totals

Table 2 showed the Principal Examiner's concurrent marks ((b)) to have been more generous than the three assistant examiners' average marks ((c)) on all bar three scripts⁵, with a mean difference of three marks, and with a similar mean difference for the amalgamated examiners' average ((d)). Table 4(a) contrasts the script totals of each amalgamated examiner A1 to A5 with the Principal Examiner's concurrent script totals, which in four cases out of five are more generous.

The relative generosity of the Principal Examiner's marking may simply reflect commonly found differences between experienced senior examiners and their less experienced assistant examiners which have not been dispelled during the standardisation process. The scale of the difference is unlikely to be connected to the introduction of the component to e-marking or indeed to the design of the present study.

⁴ An 'incestuous' feature of the data should be noted again at this point, that the examiners who have contributed to the amalgamated examiner totals, by donating (at most) one of the marks for the part questions 111+2, 113 etc., include the Principal Examiner (see para. 3.1).

⁵ Table 2 also showed that the Principal Examiner's concurrent marks were more generous than the pre-standardisation marks ((a)) on all bar four scripts, with a mean difference of three marks. However doubts about the pre-standardisation marks have been expressed in para. 3.1.

Table 4(a) Amalgamated examiners' script totals versus the PEx 'true' live marks

script	amalgamated examiner					PEx live
	A1	A2	A3	A4	A5	
1	44	52	42	45	43	46
2	51	51	55	53	49	51
3	51	48	49	47	51	53
4	35	32	42	37	41	35
5	38	36	37	26	34	35
6	35	41	38	32	42	42
7	43	45	40	47	44	45
8	31	23	27	22	20	28
9	30	28	36	29	28	36
10	30	27	33	36	34	34
11	17	23	26	30	19	27
12	36	33	27	32	33	36
13	42	36	40	38	38	43
14	27	25	30	26	28	32
15	39	47	47	41	53	52
16	30	30	34	34	18 ⁶	33
17	40	46	40	43	36	42
mean	36.41	36.65	37.82	36.35	35.94	39.41

3.3.2 Absolute mark differences in script totals

Table 4(b) gives the AMDs of the differences between the script totals of each amalgamated examiner A1 to A5 and the Principal Examiner's concurrent script totals in Table 4(a). Seven instances are emboldened where the total differs from the Principal Examiner's total by more than twice the typical grade difference of four marks (i.e. AMD > 8). There are 25 other instances where the total differs by more than one grade difference, while the remainder (53 or 62%) are within one grade, and two in five (34 or 40%) are within a half grade (2 marks). In eight cases there is complete agreement between the totals and in another thirteen the totals differ by one mark (25% combined).

The mean difference of three marks between the Principal Examiner's and the amalgamated examiners' marks has contributed substantially to the mean AMD in Table 4(b) of four marks (the equivalent of a whole grade width on the paper) and thus to marking reliability appearing to be weak.

⁶ The script total mark of 18 for script 16 is so disparate as to question the keying accuracy of one or more of the examiners contributing marks to amalgamated examiner A5.

Table 4(b) Script total AMDs by amalgamated examiner

script	amalgamated examiner					
	A1	A2	A3	A4	A5	
1	2	6	4	1	3	
2	0	0	4	2	2	
3	2	5	4	6	2	
4	0	3	7	2	6	
5	3	1	2	9	1	
6	7	1	4	10	0	
7	2	0	5	2	1	
8	3	5	1	6	8	
9	6	8	0	7	8	
10	4	7	1	2	0	
11	10	4	1	3	8	
12	0	3	9	4	3	
13	1	7	3	5	5	
14	5	7	2	6	4	
15	13	5	5	11	1	
16	3	3	1	1	15	
17	2	4	2	1	6	
mean AMD	3.71	4.06	3.24	4.59	4.29	3.98
as % of one grade width	93%	101%	81%	115%	107%	99.4%
s.d.	3.6	2.5	2.4	3.3	3.9	

3.3.3 Correlations between script totals

The coefficients of correlation between the individual amalgamated examiner and true mark totals are shown in Table 5, and between the amalgamated examiners in Table 6. Marking reliability would appear to be strong on the basis of the correlations in Table 5; they are not influenced by the relative generosity of the Principal Examiner's marks other than, as already noted, through there being an 'incestuous' feature in the data in that the Principal Examiner's marks contribute in small part to the amalgamated examiner totals. This means that the coefficients in both these tables are upper limits. A further limitation is that only 17 scripts are represented.

Table 5 Correlations between amalgamated examiner and 'true' script totals

A1	A2	A3	A4	A5
.875	.926	.891	.838	.916

Table 6 Correlations between amalgamated examiners' script totals

	A1	A2	A3	A4
A2	.872			
A3	.835	.838		
A4	.775	.845	.824	
A5	.815	.848	.852	.756

3.3.3 Comparisons with paper-based marking

A differently designed study would be needed to establish whether or not there is statistically significantly less agreement between examiners in concurrent on-screen marking of essay papers compared with post-awards, paper-based marking (as was found for GCSE English in an earlier investigation of the converse of live, paper-based marking with post-awards, on-screen marking of long answers (Fowles, 2006b)). The present investigation is concerned only with the reliability of the component total marks from on-screen marking by different examiners. Moving from a post-awards approach to an approach based on concurrent data inevitably rules out direct comparisons with paper-based marking.

Although no equivalent estimate of AMD-based or correlational marking reliability is available for paper-based marking of this particular History paper (Paper 3), evidence for Paper 1 of the same specification is available from data collected for a research investigation of online standardisation (Taylor, Chamberlain and Meadows, 2008). The marks awarded by 89 examiners to 30 post-standardisation scripts were compared with the Principal Examiner's 'true' marks. The average AMD was 2.39 out of a maximum mark of 75 (Table 3, p10), which is half the average AMD for the amalgamated marks in Table 4(b), which is 3.98 out of 60. However this cannot be viewed as a statistically significant difference given the very small numbers of scripts and examiners in the present study.

It should be noted that, because the spread of marks over the individual examiners (c) in Table 2 was greater than for the amalgamated examiners (d), e-marking AMDs from single examiners would inevitably compare more unfavourably with the (single-examiner) AMDs from paper-based marking (in the online standardisation study) than those of the amalgamated examiners.

3.4 Examiner feedback

Feedback from the examining panel was sought in a post-marking questionnaire to the History and Classical Civilisation examining panels and was also collected during the marking period by the Principal Manager with responsibility for liaising with DRS and introducing the CMI+ LFA software. For Classical Civilisation there were six assistant examiner responses to the questionnaire, but only one for History. This was perhaps because the examiners felt they had already 'had their say' in their communications with AQA, particularly the Principal Manager, who also held a post-awards meeting with the senior History examiners.

The Principal Manager has produced a very positive report on the marking, with findings that include the following:

- (a) the question paper re-numbering and the use of the new generic CMI+ answer booklet were satisfactory;
- (b) no difficulties arose from script scanning and the presentation of candidates' responses for marking (but note that a relative high proportion of scripts had to be rejected at the scanning stage and returned to paper-based marking);
- (c) choice of options and choice of questions within options has both positive and negative consequences. Because some optional questions are selected only rarely, the previous paper-based procedures meant that senior examiners did not necessarily see the complete range of optional questions. On-screen marking therefore helps in preparation of mark schemes. It also assists the Principal Examiner in selecting suitable script ranges for the Awards Meeting to consider.

The Principal Manager has made recommendations for further development that have been discussed with DRS and will guide future development of the software and procedures. It was accepted when the June 2009 trials went ahead that the software did not yet have its full functionality.

There was one particular software feature that examiners were unhappy with in their questionnaire responses, which concerned their quotas of answers to be marked. They were asked how important it was that accurate details are displayed of marking completed and marking still to do. Examiner quotas of optional questions are difficult to determine and communicate to examiners in e-marking, hence examiners were left unsure as to how many responses to any particular question they could expect or be expected to mark. This caused disquiet, as in this response from the History examiner:

“Having an easy to view summary of marking done and marking still to do is vital for a number of reasons:

- 1. planning of marking to meet deadlines and fit work around other commitments*
- 2. encouragement of knowing how much examiners are earning..... The current situation leaves examiners working in the dark – are there 4 or 44 of a question still to do?*
- 3. is it worth staying with a question to finish it or shall I move to another question for a change?”*

and this response from a GCE Classical Civilisation examiner:

“If it was possible, it would be good to have a minimum as well as maximum quota per question for each marker. It was also difficult to plan my time as there was no way of knowing how near I was to the end of available papers. It would be very good to have some sort of counter which told you how many of each question were still unmarked.”

An issue raised by examiners was noted in this response from another GCE Classical Civilisation examiner in connection with the limited provision (in CMI+ generally) to return to, and if necessary re-mark, previously marked responses, which is restricted to the immediately preceding response:

“Marking essays and keeping to a standard requires one to look back at a range of previously marked scripts. This is not possible with CMI+ LFA, so slows marking down considerably and introduces considerable doubt in arriving at marks”

To offset this uncertainty in marking LFAs the examiner requested that there be a ‘*considerable increase in model marked scripts/standardising material*’.

A senior GCE Classical Civilisation examiner drew attention to a problem with some scripts that is exacerbated in on-screen marking. His comment looks forward to the delivery of assessments online:

“It may be a pipedream, but if this is the only time that (candidates) don't word process, shouldn't their responses here be word processed as soon as this is possible? I had a dreadful headache throughout the marking period as a MAJORITY of papers were either very hard to decipher because of poor handwriting, or so badly spelt and punctuated that it took a long time to work out what they were saying. This would be bad enough with a paper copy in front of you, but felt like the last straw at times given the need to navigate backwards and forwards online.”

This examiner then summed up with a comment in relation to his e-marking that it “*requires examiners to change previous philosophy of marking, but in itself seemed to work*”.

4. DISCUSSION POINTS

4.1 The scope of the investigation

The move from paper-based marking of essays to electronic marking from screen images raises any number of issues surrounding the cognitive demands of the marking activity and their impact on the quality of on-screen marking in comparison with paper-based marking (see for example Johnson and Nádas (2009) for a discussion of factors that might affect on-screen marking consistency). The present investigation is concerned only with the reliability of the component total marks from on-screen marking by different examiners, and reference to findings from paper-based marking has been made only for the purpose of providing a context in which to gauge the acceptability of its (tentative) results. This is not to suggest that a wider approach should not be taken in future studies to help understand the cognitive processes that are involved in the change from a ‘*previous philosophy of marking*’.

4.2 The definition of ‘true’ marks

The Principal Examiner’s concurrent marks, which have been taken as the ‘true’ marks, tended to be more generous than those of the assistant examiners, and this resulted in high AMDs. They were also more generous than the pre-standardisation marks. Doubts about the consistency of the Principal Examiner’s on-screen marking therefore led to initial set of analyses which, in place of the usual hierarchical definition, used an alternative, consensus definition of true marks that gave the same status to the Principal Examiner’s marks and the assistant examiners’ marks⁷. Subsequently the analyses reverted to the hierarchical definition, because the consensus definition does not reflect AQA’s various hierarchical operational procedures in respect of the marking of the Principal Examiner and Team Leaders, nor does it conform to previous AQA marking reliability studies that have also typically used the hierarchical definition of the true mark (e.g. Fowles, 2009).

4.3 Directed marking

In category (d) the examiners have contributed in roughly equal measure to the script totals, thus recognising the benefits of segmentation. Another feature of segmented e-marking with CMI+ is the facility to direct examiners to particular questions or away from others (for example the more specialised or less popular questions that examiners in general might be less confident in marking). This suggests another way to aggregate S portion marks and generate ‘amalgamated’ script totals in future work of this kind. In the context of this History paper, each of the four concurrent examiners’ marks could be directed to each of the four question parts in turn, giving 24 permutations and thus 24 more sets of amalgamated total marks for comparison. This would be the more valid ‘amalgamating’ approach to use in estimating reliability where specialisation of marking was widely adopted.

4.4 Procedural issues

An inadequate quantity of research data was collected because of procedural difficulties, most of which can be avoided in future use of the S portion approach to collecting concurrent marking reliability data. The pre-standardisation marking will inevitably be to a tight time schedule, but the time could be increased by making the S portions available to the Principal Examiner for

⁷ The more positive but still very tentative results using this approach to defining true marks were reported to the meeting of the AQA/DRS Steering Group meeting on 30 September 2009.

home marking. The marking needs to work to a tightly drawn and final mark scheme, and to allow for unforeseen wastage in the selected sample. With the S portion procedure running more smoothly there is every reason to expect that a sufficient sample of scripts would be available in using the approach in future series.

There are some other aspects of the e-marking procedure that are likely to have affected the quality of the e-marking in the trial History paper.

- The Principal Examiner requested DRS to link certain part questions because he expected examiners to prefer the continuity. The linking certainly assisted the analysis, by effectively reducing the number of responses to the paper per candidate from seven to just four questions. However it may have had a detrimental effect on the quality of the marking, because the examiners reported finding it difficult to keep the relatively long mark schemes for the separate linked questions in mind. The full extent of benefits associated with segmentation in e-marking, with marking focussed on single questions, was lost. The Principal Examiner and Team Leaders have requested that this linkage be discontinued in future series.
- The facility to return to the previously marked response was not available for this first trial with the LFA software. This was also a source of (temporary) examiner anxiety when CMI+ was first introduced, but is expected to be rectified for the next series.

4.5 Viewing previously marked clips

In addition to revisiting the single previously marked response, LFA examiners have expressed a wish to be able to retrieve a much larger pool of marked responses. Whereas examiners using CMI+ are occasionally aware they have made an error and need to go back to correct it (especially on a very easy or a very difficult short answer question where they tend to develop a mark predisposition), with LFA marking there is an additional need to be able to review earlier marking for consistency, and if necessary to amend it. This is most keenly felt in a settling-in phase when starting on a new LFA question. Although increasing the size of the pool of responses that can be re-marked cannot be contemplated at this stage, a new facility is planned for the LFA software which will allow a number of earlier clips and the marks awarded to be stored and viewed (but not changed).

A suggestion for improving the quality of e-marking in subsequent series is to routinely start the marking of each question with a standard set of S portions as qualification items. This would also help alleviate examiner anxiety about early marking inconsistency. However there are operational and other considerations that would need costing and discussion within AQA before this feature could be implemented. A more straightforward alternative is a pre-populated pool of marked responses available for examiners to refer to from the start of their marking, using the new facility to store and view clips.

5. CONCLUSION

The Principal Examiner's concurrent marks tended to be more generous than both the assistant examiners' marks and the amalgamated examiners' marks and this contributed to apparently weak marking reliability as gauged by the AMDs between the Principal Examiner's marks and the amalgamated examiners' marks (the mean AMD in Table 4(b) was four marks, the equivalent of a whole grade width on the paper). Rather greater variability in the AMDs would be found had individual examiners' script totals been used, although it is argued that the pooled data approach is more appropriate in that it reflects the segmented nature of electronic marking

rather than the whole paper approach of paper-based marking. Marking reliability appeared much stronger as gauged by the correlations between the Principal Examiner's and the amalgamated examiners' marks (Table 5). However, it is acknowledged that difficulties in completing the marking of the S portions as intended meant that an adequate quantity of research data could not be collected and the analyses have been based on a very small number of examiners and scripts. The very limited evidence from the e-marking of this LFA, essay style History paper is thus too tentative to determine, from either variability in script totals or from the correlational relationships between script totals, whether or not marking reliability is at expected levels when marks from different examiners are pooled to give those script totals. More substantial evidence from a variety of papers should be sought from future series.

Although flawed in operation in this first trial, the use of the S portion approach has the potential to provide good quality, concurrent data for investigating the reliability of on-screen marking from a sample of scripts. The method capitalises on the facility in electronic marking to gain concurrent marking data for the same scripts from every member of a team of examiners, without their awareness of which scripts are being used to generate the data (operationally unobtainable in paper-based marking). Concurrent marking data would be extremely valuable for a credibility check on the suitability of electronic marking for the variety of long form answer components that it is intended should be electronically marked in future series.

The recommendation of this project is therefore that AQA should continue to explore the reliability of on-screen marking of long form answers, as more examiners and more components are brought into the process, using S portions to collect concurrent marking reliability data (the next opportunity is the marking of January 2010 GCE units).

Dee Fowles
Research and Policy Analysis Department
November 2009

Z:\e-marking reliability LFAs 2009\Reliability of LFA e-marking in GCSE History.doc

REFERENCES

- Fowles, D. (2006a) *How reliable is the marking in GCSE English?* AQA Research Committee paper, RPA_06_DF_RP_020.
- Fowles, D. (2006b) *How well does marking in GCSE English transfer to marking using CMI+ with annotation?* AQA Research Committee paper, RPA_06_DF_RP_047.
- Fowles, D. (2009) *How reliable is the marking in GCSE English?* *English in Education*, v43, Issue 1, p50-67.
- Johnson, M and Nádas, R. (2009) Investigation into marker reliability and some qualitative aspects of on-screen essay marking. Cambridge Assessment's *Research Matters*, Issue 8, June 2009, p2-7.
- Meadows, M.L. and Billington, L. (2005) *A Review of the Literature on Marking Reliability*. AQA Research Committee paper, RPA_05_MM_RP_005 (RC/304).
- Taylor, R., Chamberlain, S. and Meadows, M. (2008) *Comparing the effects of on-line and face-to-face training on marking reliability*. AQA Research Committee paper RPA_08_RT_RP_053.