Online standardisation trial, Winter 2008: Evaluation of examiner performance and examiner satisfaction

Lucy Billington & Chloe Davenport

SUMMARY

In January 2008, online standardisation was trialled for the second time across six components. Half of the components were marked conventionally and half at item level and onscreen using CMI+. For the conventionally marked components, examiners' quality of marking was explored using first and second phase sample data from the January 2007 and January 2008 series. For two out of three conventionally marked components, there was no significant difference in terms of absolute mark differences between the paper-based second phase samples completed in January 2007 (following face-to-face standardisation) and January 2008 (following online standardisation). Larger absolute mark differences were observed for online first and second phase samples (following online standardisation) than for paper-based samples (following face-to-face standardisation). This finding is consistent with those from the first trial in June 2007, and is believed to be a result of procedural differences in the way that the two types of samples are selected and completed.

The GCE Sport and Physical Education component, PED4 was included in the first online standardisation trial in June 2007. The findings for this component were exceptional, in that mode of standardisation appeared to have impacted upon the accuracy of the examiners' second phase sample paper-based marking. In addition, their online second phase sample marking was found to have deteriorated since the previous trial. The mark scheme for the January 2008 paper may have been more complex and translated less well to online standardisation. Also an increased workload, due to participation in a research exercise and being standardised online for an additional component (PED1 was introduced to online standardisation in the second trial), offer potential explanations for the examiners' reduced marking reliability.

Limited data was available for the three components marked by CMI+ but it appeared that online standardisation did not impact upon the examiners' ability to mark seed items. The findings for the January 2008 data are discussed in the light of the recent findings of other online standardisation research initiatives e.g. an experimental study conducted with GCSE History examiners. Whilst there is substantial quantitative evidence to support the extension of online standardisation to other components, qualitative findings from evaluation questionnaires administered to participants suggest that examiners still have concerns regarding the transition from face-to-face to online standardisation. Some of these concerns stem from technical difficulties experienced with the online resource during the second trial.

BACKGROUND

In January 2008 the Assessment and Qualifications Alliance (AQA) trialled its online standardisation facility for the second time. The trial included a total of six components (see Table 1) and involved 74 examiners. Only two components (3820/3 and PED4) were included in the previous online standardisation trial, which took place in June 2007. January 2008 was the first time that online standardisation had been combined with CMI+. CMI+ is a software package that enables examiners to mark scanned images of scripts, using their home PC. Examiners mark scripts directly on-screen at item level.

Table 1: Components included in the January 2008 online standardisation trial.

Qualification	Code	Title	CMI+ Jan 08
GCSE Health & Social Care	*3820/3	Understanding Personal Development and Relationships	×
GCE Sport & Physical Education	PED1	Physiological and Psychological Factors	×
GCE Sport & Physical Education	*PED4	Physiological, Biomechanical and Psychological Factors	×
GCE Computing	CPT2	Principles of Hardware, Software and Applications	✓
GCE Computing	CPT5	Advanced Systems Development	✓
GCE Geography B	GGB1	The Dynamics of Change	✓

*Components/units included in the June 2007 trial.

Chamberlain (2007) reported on the findings of the first live online standardisation trial in summer 2007. Examiners for 3820/3 and PED4 completed the standardisation process and their first and second phase samples online. These samples of scripts were pre-selected by the Principal Examiner (PEx) and had pre-determined marks. Examiners marked these scripts online free from any marks and annotations. Each examiner also submitted a conventional (paper-based), self-selected second phase sample which was over-marked by Team Leaders and on which any examiner adjustments were based.

Chamberlain found that marking first and second phase samples online (following online standardisation) resulted in larger absolute mark differences than when the samples were marked on paper in the previous year (following conventional standardisation). These larger mark differences were likely to be a product of crucial distinctions between the two marking environments. There is evidence, for example, that a comparison of the marks awarded by examiners with the pre-determined marks of the PEx leads to greater mark differences than when Team Leaders are able to be influenced by the examiners' marking (Baird and Meadows, Under review). Indeed, much smaller mark differences were observed for the self-selected second phase sample which was marked on paper following online standardisation in June 2007. Furthermore, a consideration of the total script scores awarded by the PEx against the mean score awarded by the examiners for the online second phase sample showed a high

degree of agreement. Thus, whilst large differences existed at item level for online marking, it was reassuring that at script level there were few significant differences.

As part of the June 2007 trial, examiners were asked to complete a short evaluation questionnaire to gauge their perceptions of the process and their feedback on the quality of the online standardisation facility. Whilst examiners acknowledged a number of benefits to online standardisation (e.g. quality and speed of feedback, being able to revisit materials), they expressed concern over the loss of face-to-face meetings and the opportunity to engage with the examiner community.

This paper aims to replicate the analysis performed by Chamberlain for the January 2008 trial data. Section 1 is concerned with exploring examiners' quality of marking. Analyses are presented separately for those components marked conventionally and those marked electronically in January 2008. Examiners for 3820/3, PED4, and PED1 completed their first phase samples online and two second phase samples (online and paper) in January 2008. For the first phase sample data, comparisons are made (across items) between matched examiners' performances in January 2007 (when they received face-to-face standardisation and marked paper samples) and January 2008 (following online standardisation and using online samples).

The analyses for the second phase sample data entails two comparisons: 1) a comparison of the impact of mode of standardisation on paper-based marking (January 2007 paper samples versus January 2008 paper samples); and 2) a comparison between online and paper marking following online standardisation in January 2008. For those components that were included in the first trial in June 2007, comparisons are also made between online first and second phase samples to assess whether the examiners' online marking has improved with greater experience of the online system. Since examiner adjustments are made at script level, total script scores are also explored for the online second phase samples.

Examiners for CPT2, CPT5 and GGB1 marked their live allocations on screen using CMI+ in January 2008. Examiner performance for online marking was monitored throughout the marking period using qualification and seed items, which had been previously marked by the PEx. Examiner level information pertaining to item seeds is used to explore the impact of online standardisation on examiners' quality of marking.

Section 2 summaries the outcomes of an evaluation questionnaire administered to all examiners shortly after the January 2008 trial. The questionnaire was similar to that completed by participants following the June trial, with some amendments to take into account whether examiners were new to online standardisation or had participated in the previous trial. Before exploring examiners' quality of marking, it is necessary to highlight some of the limitations of conducting comparative work with online and paper-based sample data.

Some limitations of comparing online and paper-based samples

The limitations of drawing comparisons between online and paper-based samples are well-documented by Chamberlain (2007). The most generic include the limited amount of first and second phase sample data available for such analyses; and examiners' greater familiarity with conventional standardisation procedures. Most importantly, Chamberlain (2007) notes that the 'conventional' or 'paper-based' and 'online' labels merely *"serve as umbrella terms for a range of important procedural differences between the marking environments"* (p.18).

The procedure used to select scripts for inclusion in the first and second phase sample represents the most striking difference between the two methods. Under the conventional method, examiners self-select from their marking allocation the scripts they wish to be included in their first and second phase samples. A selection of these is then over-marked by Team Leaders. Importantly, the Team Leader is able to see the marks and annotations of the first examiner. In contrast, the online standardisation system uses common scripts for both the first and second phase sample. These scripts are selected and marked by the PEx, and are presented to examiners in a clean state (free from any marks or annotations). Thus, the procedure for selecting online second phase samples prevents examiners from selecting their 'best' scripts, and avoids any bias caused by the presence of the first examiner's marks and annotations. Whilst the online procedure may be more robust in terms of monitoring examiner performance[1], it is somewhat inevitable that it will produce larger absolute mark differences than the conventional method.

Furthermore, in conventional (face-to-face) standardisation training Team Leaders have the opportunity to influence examiners' interpretation of the mark scheme, since they facilitate the standardisation meeting and over-mark the paper-based first and second phase samples. It seems likely, therefore, that there will be closer agreement between the marks awarded by Team Leaders and examiners, than those awarded by the PEx and examiners (the benchmark against which the quality of examiners' online sample marking is judged). Indeed, Meadows and Taylor (2008) have identified the removal of Team Leader influence as a potential benefit of online standardisation – *"examiners' standardisation is standardised"* (p.9).

To evaluate the impact of online standardisation on first phase sample marking, it was only possible to compare online and paper-based samples. The limitations outlined here, should be borne in mind when drawing conclusions about the online resource or the quality of first phase sample marking following online standardisation. In the case of the second phase sample data, a comparison could be made between paper-based second phase samples completed in January 2007 (following face-to-face standardisation) and January 2008 (following online standardisation). This like-with-like comparison should be regarded as most telling in terms of the impact on online standardisation on examiners' quality of marking. Having said this, a comparison between online and paper-based second phase samples following online standardisation in January 2008 has been included. It was thought that such a comparison may be helpful in considering the implications of replacing paper-based second phase samples with online second phase samples in the future.

SECTION 1: QUALITY OF MARKING

Non CMI+ components

Comparing examiners' First Phase Sample performances

The analysis below compares examiner first phase sample marking performances (of 10 scripts per sample) in January 2007 (on paper) and January 2008 (online). The data are for matched

[1] It should be noted, however, that online sampling is not without its pitfalls. The use of common scripts means that judgements made about examiners' quality of marking are not based on live marking. Furthermore, online sampling has resource implications as time and money are spent marking common rather than live scripts.

examiners only. It should be noted that usable data were not obtainable for all participants. There were 5 examiners common to January 2007 and 2008 for 3820/3, 18 examiners for PED1, and 8 examiners for PED4. Due to the small samples sizes involved, the findings should be treated with caution.

For each component the examiners' absolute mark differences (from the marks awarded by the PEx) have been averaged over each question (8 for 3820/3, 5 for PED1, and 5 for PED4) and each script (10), producing for each examiner a mean absolute mark difference (AMD). A final grand mean absolute raw mark difference across all examiners was calculated per component and converted to a percentage of the maximum raw score.

Table 2 shows the differences between January 2007 and 2008 for each component. The mean absolute mark differences between the January series are fairly large for each of the components, albeit less so for PED1. The differences are all statistically significant, suggesting that the observed variation between January 2007 and 2008 is not due to chance alone. Indeed, the effect size (using Cohen's *D*) reported for each component is considered to be 'large'. Cohen (1992) suggests that 0.2 is indicative of a small effect, 0.5 a medium and 0.8 a large effect size. These finding replicate those of the first trial (see Chamberlain, 2007).

Additional analysis for those examiners that participated in the June 2007 trial showed a non-significant difference between their online first phase sample marking in June 2007 and January 2008 (3820/3 = $t_{(4)}$=0.50, p=0.641, PED4= $t_{(6)}$=0.90, p=0.402). This finding suggests that the examiners' online first phase sample marking has remained stable between trials (it has not improved nor deteriorated).

Table 2: Examiners' first phase sample absolute mark differences.

		GCSE 3820/3 % scores (raw max = 95 in 2007; 98 in 2008)		GCE PED1 % scores (raw max excl QWC= 72)[2]		GCE PED4 % scores (raw max excl QWC = 60[3])	
		2007 paper	**2008 online**	**2007 paper**	**2008 online**	**2007 paper**	**2008 online**
Number of examiners		5		18		8	
Minimum mean AMD		0.63	4.59	1.11	3.33	0.83	5.33
Maximum mean AMD		1.89	7.76	5.83	8.06	3.17	9.67
Grand mean AMD (% of max raw score)		1.41	6.33	3.17	4.95	1.60	6.60
Std Deviation		0.50	1.16	1.18	1.00	0.79	1.32
95% CIs	**Lower**	0.79	4.89	2.58	4.45	0.95	5.50
	Upper	2.03	7.76	3.76	5.44	2.27	7.71
Median AMD		1.37	6.53	2.99	4.72	1.25	6.33
***t*-test results**		$t_{(4)}$=8.16, p=0.001; Cohen's D =5.51		$t_{(17)}$= 5.34 p<0.001; Cohen's D =1.63		$t_{(7)}$=8.14 p<0.001; Cohen's D =4.60	

Comparing examiners' Second Phase Sample performances

The second phase sample data (consisting of 15 scripts per sample) from the January trial were also explored across items, using the same methodology as above. Examiners completed two second phase samples in January 2008 (paper-based and online). In the first instance, a comparison is made between the January 2007 and January 2008 paper-based samples. This like-with-like comparison is most useful in terms of exploring the impact of mode of standardisation on examiners' quality of marking. A second comparison is made between the paper-based and online second phase samples from January 2008, following online standardisation. As discussed above, any observed differences in the accuracy of examiners' marking of paper-based and online second phase samples is likely to be the result of procedures used to select and complete the two sample types. Such a comparison, however, may be informative in terms of amendments needed to business rules should online second phase samples replace paper-based second phase samples as the basis for examiner adjustments in the future. The maximum possible number of examiners has been included in each comparison[4].

[2]For PED1, 3 marks were awarded for Quality of Written Communication (QWC). QWC was not monitored during the online standardisation process, and has consequently been omitted from all analyses for this component.

[3]For PED4, 4 marks were awarded for QWC. QWC was not monitored during the online standardisation process, and has consequently been omitted from all analyses for this component.

[4] In other words, it was not necessary for examiners to have usable data for all three second phase samples to be included in the analysis.

Table 3 shows that for all three conventionally marked components examiners exhibited a larger grand mean AMD (as a percentage of the maximum mark) for the paper-based second phase sample in January 2008 (following online standardisation) than the paper-based sample in January 2007 (following conventional standardisation). However, the results of the paired samples *t*-tests suggest that these differences were non-significant for two out of the three components (3820/3 and PED1). In other words, mode of standardisation did not significantly impact upon 3820/3 and PED1 examiners' marking accuracy of paper-based second phase samples. The reported effect sizes suggest that online standardisation had a 'small' effect on PED1 examiners' quality of marking (Cohen's *D*=0.28) and a 'large' effect on 3820/3 and PED4 examiners' paper-based marking (Cohen's *D*=1.06 and 1.64, respectively).

In contrast to the current findings, online standardisation in June 2007 appeared to have impacted upon the 3820/3 examiners' conventional second phase sample marking (see Chamberlain, 2007). The disparity in findings between years might indicate an improvement in 2008 due to examiner's familiarisation with the online standardisation facility, or the January 2008 paper being easier to mark. Having said this, a comparison of the June 2007 and January 2008 online second phase samples revealed a non-significant difference in percentage mean AMD ($t_{(5)}$= 0.050 p=0.963). Thus, whilst for 3820/3 examiners mode of standardisation had a diminished effect on their paper-based marking, the quality of their online marking remained similar to that of the first trial.

In June 2007 online standardisation did not significantly impact upon PED4 examiners' paper based marking of second phase samples (see Chamberlain, 2007). However, in January 2008 a significant difference was found between the two paper samples from the January series. PED4 examiners marking of online second phase samples was also found to have deteriorated between the two trials ($t_{(5)}$=5.88, p<0.001). Possible explanations for the worsening of PED4 examiners' second phase sample marking since June 2007 might include increased complexity of the paper and mark scheme and/or a change in the composition of the online second phase sample. The PEx may have selected a more 'problematic' sample of common second phase sample scripts compared to June 2007. In addition, a number of PED4 examiners were also marking for PED1 in January 2008 and completed a third second phase sample (the outcomes of which are discussed elsewhere - see Billington, In preparation) as part of a research exercise. The PED4 examiners' quality of second phase sample marking may have suffered as a consequence of an increased workload.

Table 3: Examiners' second phase sample absolute mark differences (paper 2007 versus paper 2008).

		GCSE 3820/3		GCE PED1		GCE PED4	
		% scores (raw max = 95 in 2007; 98 in 2008)		% scores (raw max excl QWC= 72)		% scores (raw max excl QWC = 60)	
		2007 paper	2008 paper	2007 paper	2008 paper	2007 paper	2008 paper
Number of examiners		7		20		12	
Minimum mean AMD		0.42	0.61	0.56	0.65	0.44	2.00
Maximum mean AMD		3.09	5.99	4.17	7.78	2.56	6.22
Grand mean AMD (% of max raw score)		1.25	2.81	2.23	2.65	1.45	3.18
Std Deviation		0.92	1.87	1.15	1.78	0.67	1.33
95% CIs	**Lower**	0.40	1.08	1.70	1.81	1.03	2.33
	Upper	2.10	4.54	2.77	3.48	1.88	4.02
Median AMD		1.12	2.52	2.22	2.50	1.39	2.72
***t*-test results**		$t_{(6)}$=2.11, p=0.080; Cohen's D =1.06		$t_{(19)}$=1.00, p=0.332; Cohen's D =0.28		$t_{(11)}$=4.53, p=0.001; Cohen's D =1.64	

Table 4 compares matched examiner performances for the paper-based and online second phase samples in January 2008, following online standardisation. As expected, larger grand mean AMDs were observed for the online than the paper-based second phase samples. Moreover, in each case, a significant difference was found between the examiners' marking of paper and online second phase samples in January 2008 (as shown by the results of the paired samples *t*-tests). The values of Cohen's *D* show that online standardisation had a 'large' effect on online marking for all three components, but that this effect was most marked for PED4 examiners (Cohen's D=4.09).

Second phase sample forms are the primary evidence used in deciding the outcome of the adjustments procedure. In reviewing, second phase sample forms staff must take into account the marking 'tolerance'. The tolerance is 5 *per cent* (rounded upwards) of the maximum mark for a given component, and is applied at total script level. For example, if an examiner's marking is not in complete accordance with the senior examiner's marking, but the differences are within tolerance the decision will be 'no adjustment'. Although, the differences reported in Table 4 are across items rather than at total script level, they imply that larger tolerances may be required should online second phase samples become standard practice.

Table 4: Examiners' second phase sample absolute mark differences (paper 2008 versus online 2008).

		GCSE 3820/3		GCE PED1		GCE PED4	
		% scores (raw max = 95 in 2007; 98 in 2008)		% scores (raw max excl QWC= 72)		% scores (raw max excl QWC = 60)	
		2008 paper	2008 online	2008 paper	2008 online	2008 paper	2008 online
Number of examiners		6		21		12	
Minimum mean AMD		0.61	6.19	0.65	4.44	2.33	6.78
Maximum mean AMD		5.99	8.16	7.78	7.13	5.56	9.78
Grand mean AMD (% of max raw score)		3.31	7.34	3.32	6.13	3.55	8.10
Std Deviation		1.78	0.90	1.64	0.63	1.16	1.06
95% CIs	**Lower**	1.44	6.39	2.57	5.85	2.81	7.43
	Upper	5.18	8.28	4.06	6.42	4.28	8.77
Median AMD		3.30	7.65	2.87	6.02	3.11	7.61
***t*-test results**		$t_{(5)}$=7.33, p=0.001; Cohen's D =2.86		$t_{(20)}$=7.17 , p<0.001; Cohen's D =2.26		$t_{(11)}$= 11.05, p<0.001; Cohen's D =4.09	

Total script scores

The findings of the analyses above have focussed upon absolute mark differences at item level to ensure that quality of marking is assessed at the most precise level of measurement. It is arguable, however, that absolute mark differences in total script scores awarded by the PEx and the examiners are of greater importance. For conventionally marked components, decisions regarding examiner adjustments are made at the level of the total script.

Charts 1 to 3 show the total script scores (as a percentage of the maximum mark) awarded by the PEx and the examiners as a group for the online second phase sample for each non-CMI+ component. The mean examiner script scores were most closely aligned with those of the PEx for PED1 (Chart 2), which is surprising given that these examiners were experiencing online standardisation for the first time. The grand mean mark difference across scripts was 3.7% for PED4, 2.8% for 3820/3, and 1.3% for PED1. The maximum mean mark difference was 9.7% for 3820/3 examiners (script 6, almost 10 raw marks), 7.6% for PED4 (script 9, almost 5 raw marks), and 3.2% for PED1 (script 15, 2 raw marks). These scripts may have been selected as being particularly 'problematic' by the PEx in order to test the examiner's understanding of the mark scheme.

The current examiner adjustment process, which uses paper-second phase samples, dictates that where a difference of twice the tolerance (more than 10%) between the senior examiner's mark and examiners mark is recorded, the senior examiner's mark should be taken as the final (unadjustable) mark. The maximum mean mark differences reported for each of the conventionally marked components in the January 2008 trial provide some encouragement that large AMDs at item level (see Table 4) may not necessary translate to unadjustable marks at the total script level.

In June 2007, Chamberlain (2007) reported grand mean mark differences of 1.6% and 2.1% for PED4 and 3820/3, respectively. It is unclear why the quality of the examiners marking at script level for these components should have decreased with a second iteration of online standardisation – it may be a consequence of changes in the nature of the question paper/mark scheme, or due to the scripts included in the second phase sample. As mentioned above, PED4 examiners' workload increased in January 2008 and this may have affected their online marking. The examiners were also aware that the online second phase sample was not being used as the basis of examiner adjustments, or for the completion of performance records. They may have, therefore (quite rightly), prioritised their marking of the paper second phase sample and live candidates' scripts.

Chart 1: Comparing Principal Examiner and Assistant Examiners' total script scores, 3820/3 common online second phase sample scripts.

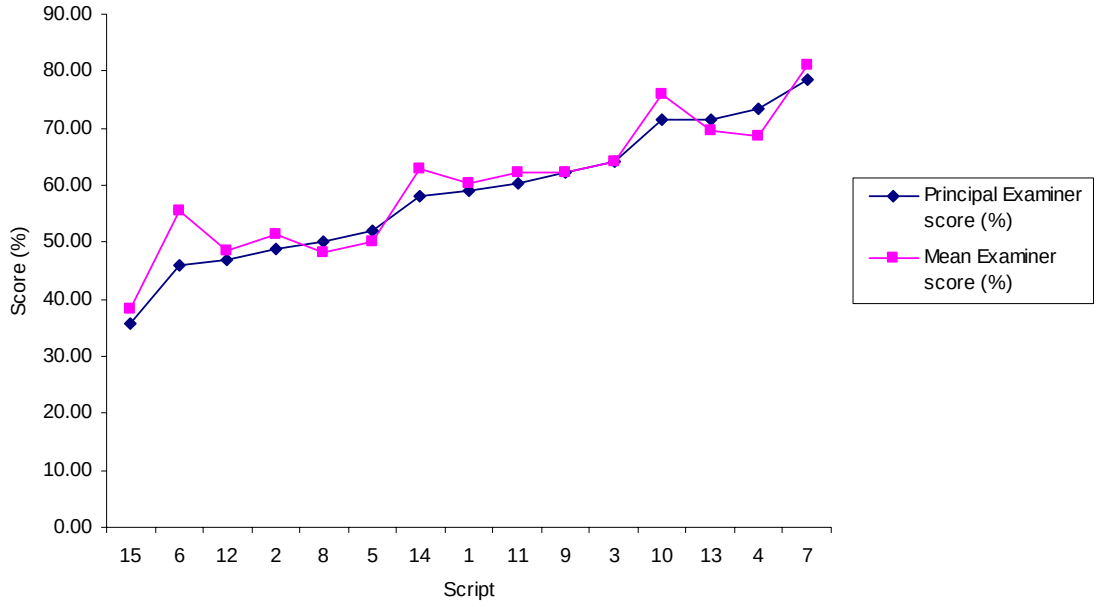


Chart 2: Comparing Principal Examiner and Assistant Examiners' total script scores, PED1 common online second phase sample scripts.

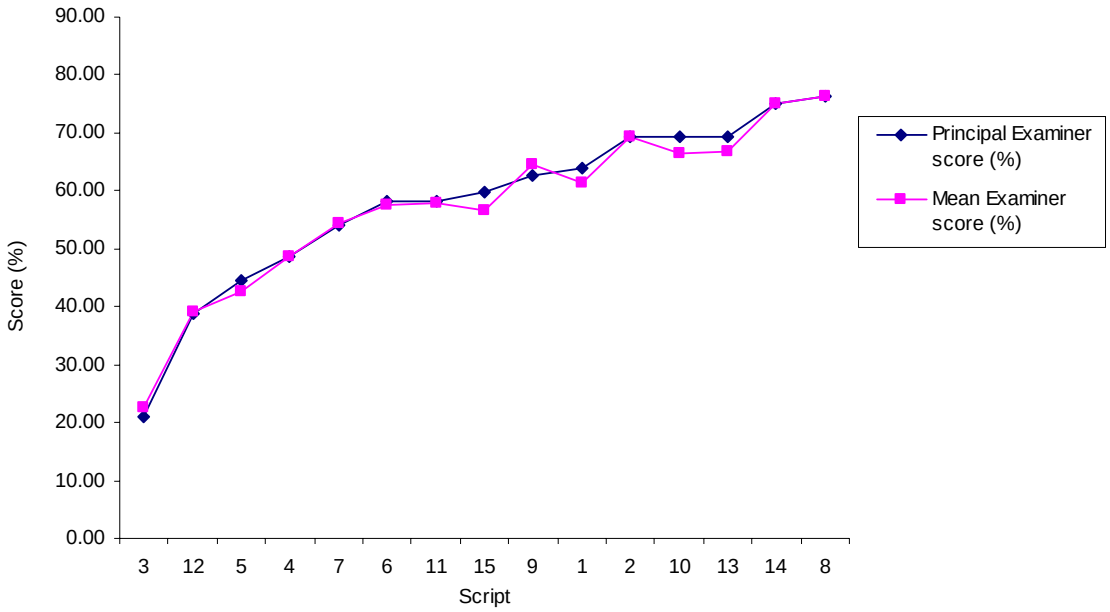
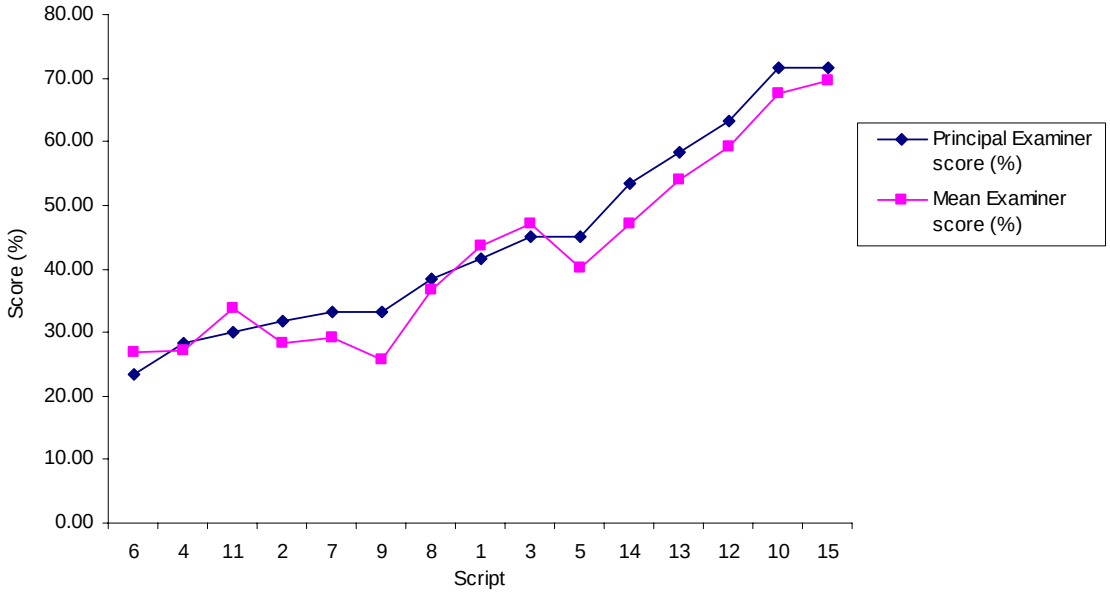


Chart 3: Comparing Principal Examiner and Assistant Examiners' total script scores, PED4 common online second phase sample scripts.



CMI+ components

CMI+ is an electronic marker facility that enables examiners to view and mark digitally imaged scripts. There are three different methods of marking on CMI+; auto, general and expert. Candidates' responses are segmented accordingly. Items designated for automatic marking are double keyed at a marking bureau and run against an agreed marking key. General and expert items are marked by examiners online, item-by-item. Examiners for CMI+ components are not required to complete first and second phase samples for over-marking by Team Leaders. Instead, throughout the marking period examiner performance is monitored using qualification and seeded items. Qualification items and seeds are selected and marked by the

PEx in advance of the marking period. To 'qualify' to mark examiners must mark eight out of ten qualification items accurately. Thereafter, a pair of seed items is planted every forty items. If the examiner fails to mark either seed item accurately they are stopped from marking the item until they have received further training. When marking is complete Team Leaders are supplied with a "Quality of Marking by Marker Report", which informs the completion of performance records of examiners in their team.

Tables 6 to 8 summarise the most pertinent figures taken from the Quality of Marking by Marker Reports for those examiners that marked the three CMI+ components in the January 2008 trial. Both CPT2 and GGB1 were marked by CMI+ for the first time in June 2007. Thus, comparisons are made between examiners performances (in terms of the % of seeds failed) in June 2007 and January 2008. CPT5 was marked by CMI+ for the first time in January 2008, preventing any comparisons to be drawn with the examiners' performance in previous series. Information pertaining to CPT5 is included for completeness. Whilst the small number of examiners for which data were available prevents generalisations, the data are largely unexceptional. For CPT2 there are no notable fluctuations in the percentage of seeds failed for the examiners involved in the trial, following online standardisation. In January 2008, CPT5 examiners tended to accurately mark a greater proportion of seeded items than their CPT2 counterparts (but not markedly so). For GGB1, there was a tendency for the percentage of seeds failed to decline following online standardisation (examiner 3 in Table 8, for example). The findings of paired-samples *t*-tests confirmed that there was no significant difference in the percentage of seeds failed by CPT2 ($t_{(4)}$=0.43, p=0.690) and GGB1 ($t_{(4)}$=1.74, p=0.157) examiners in June 2007 and January 2008.

Table 6: Extracts of the Quality of Marking by Marker Report for CPT2, June 2007 and January 2008.

	June 2007					January 2008				
Examiner	**Number Marked**	**Number seeds**	**Number failed**	**% Failed**	**Number of times stopped**	**Number Marked**	**Number seeds**	**Number failed**	**% Failed**	**Number of times stopped**
1	12582	1330	139	10.45	13	5743	627	64	10.21	4
2	11804	1255	159	12.67	17	5023	705	78	11.06	8
3	11364	1133	138	12.18	12	6148	528	66	12.50	6
4	-	-	-	-	-	5163	711	75	10.55	12
5	12146	1221	106	8.68	10	5725	659	69	10.47	5
6	12931	1218	120	9.85	13	5902	647	70	10.82	10

Table 7: An extract of the Quality of Marking by Marker Report for CPT5, January 2008.

	January 2008				
Examiner	**Number Marked**	**Number seeds**	**Number failed**	**% Failed**	**Number of times stopped**
1	5581	38	38	7.32	2
2	4792	536	48	8.96	5
3	5385	652	50	7.67	4
4	4569	612	45	7.35	4

Table 8: Extracts of the Quality of Marking by Marker Report for GGB1, June 2007 and January 2008.

	June 2007					January 2008				
Examiner	**Number Marked**	**Number seeds**	**Number failed**	**% Failed**	**Number of times stopped**	**Number Marked**	**Number seeds**	**Number failed**	**% Failed**	**Number of times stopped**
1	3030	702	64	9.1	5	2174	615	44	7.2	2
2	2543	659	45	6.8	0	1562	331	28	8.5	1
3	3624	494	77	15.6	5	2882	337	36	10.7	1
4	2912	725	54	7.4	3	2330	633	39	6.2	2
5	3252	560	43	7.7	1	2135	319	14	4.4	1

SECTION 2: EXAMINERS' RESPONSES TO THE EVALUATION QUESTIONNAIRE

This section of the report identifies the main themes of the examiners' responses to AQA via the evaluation questionnaire. Examiners were asked to complete the questionnaire as soon as possible after having received online standardisation to ensure their views were fresh from their experiences. One of two questionnaires was administered to participants, depending upon whether the examiner had taken part in the first trial in June 2007. Examiners new to online standardisation (PED1, CPT2, CPT5, and GGB1) received an identical questionnaire to that completed by examiners following the June 2007 trial. Examiners undergoing online standardisation for the second time (3820/3, PED4) received a slightly edited questionnaire to avoid repetition and to take into account their greater familiarity with the system.

The response rate was higher than in June 2007. The evaluation questionnaire following the first trial had an overall response rate of 69.6 *per cent,* which was poorer than had been expected given that examiners were informed the questionnaire was integral to the development of the resource. In January 2008, 64 of the 74 examiners (86.49%) taking part completed and returned their questionnaires. The higher response rate is thought to be attributable to the questionnaire being administered via email rather than by post. There was a slightly higher response rate for the group of examiners taking part in the trial for the first time than those taking part for the second time (89.13% compared to 82.14%). It may be the case that those who responded had a particular motivation for doing so.

Theme 1: Benefits of the system

Only those examiners that were participating in the trial for the first time were asked about the perceived benefits of completing standardisation online. As in June 2007, examiners reported the ability to complete online standardisation at their own pace as the most beneficial aspect of the system. Compared to the June 2007 examiners, the January 2008 examiners indicated greater appreciation for the more practical aspects of completing standardisation remotely e.g. completing standardisation at home or another place of choice, completing standardisation at a convenient time, and not having to take a day out of school. They were less appreciative of benefits of the system such as being able to revisit online materials (the second most popular benefit in June 2007) and raising a query with the PEx. It seems likely that 'contacting the Principal Examiner when I have a query' was the most poorly scoring item due to technical difficulties with the messaging service in January 2008.

Table 9: Benefits of online standardisation (percentage of examiners agreeing).

	Agree (%)	Total N
Completing standardisation at my own pace	61.0	41
Completing standardisation from home or another place of my choice	56.1	41
Being able to fit standardisation in around my home, family or work responsibilities	53.7	41
Completing standardisation at a time that suits me	43.9	41
Not having to travel to the AQA office/meeting venue	41.5	41
Being able to revisit the online standardisation materials when, and as often, as I need	39.0	41
Not having to take a day out of school	34.1	41
Completing standardisation on my own without distractions	29.3	41
Contacting the Principal Examiner when I have a query	24.4	41

It was commented that the system used for online standardisation was extremely slow. It is likely that frustration experienced by the examiners negatively contributed to their perceptions of the benefits of the online standardisation resource:

"I know my team found online standardisation extremely time consuming and they were very frustrated with the speed of loading and the like." (PED1)

"The time taken for parts of the process is a little disappointing as immediate does not always mean immediate." (PED1)

Theme 2: Quality of Marking

Two thirds of examiners completing online standardisation for the first time reported a decrease in their depth of understanding of the mark scheme. Examiners also reported a negative change in their preparedness for marking and ability to internalise the mark scheme (61.1 *per cent* and 48.6 *per cent*, respectively; Table 10). This finding echoes the views of the June 2007 examiners regarding the impact of online standardisation on their quality of marking.

In terms of marking aspects that examiners felt had been positively affected by completing standardisation online, a quarter of examiners reported a positive change in their awareness of the strengths and weaknesses in their marking. Over fourteen *per cen*t of examiners felt their ability to internalise the mark scheme and their concentration on the mark scheme had improved. Whilst these findings provide some encouragement, the percentage of examiners reporting a positive change in these three aspects was lower than in June 2007 (38.5% for awareness of strengths and weaknesses in marking, 15.4% for internalising the mark scheme, and 30.8% for concentration on the mark scheme). In addition, in June 2007, approximately a third of examiners reported that feedback had positively impacted upon their marking. This figure was only 11.4% in January 2008, and is likely to be attributable to technical problems with the messaging facility used by examiners to liaise with their Team Leader.

Table 10: The impact of online standardisation on first time examiners' quality of marking (percentage of examiners).

	Positive change	No change	Negative change	Total N
Awareness of the strengths and weaknesses in my marking	25.0	41.7	33.3	36
Internalising the mark scheme	14.3	37.1	48.6	35
Concentration on the mark scheme	14.3	51.4	34.3	35
Preparedness for marking	13.9	25.0	61.1	36
The impact of feedback on the quality of my marking	11.4	62.9	25.7	35
Overall quality of my marking	2.9	60.0	37.1	35
Depth of understanding of the mark scheme	0.0	33.3	66.7	36

Those examiners undertaking online standardisation for the second time were asked how they thought the quality of their marking had changed since the first trial. The majority of examiners reported that the quality of their marking had remained unchanged with greater experience of the online system (Table 11). Worryingly, 39.1% of examiners reported that, compared to the June 2007 trial, their depth of understanding of the mark scheme had changed for the worse. Approximately, 30 *per cent* also reported a negative change in the quality of their overall marking and their preparedness for marking. Such findings could be attributable, in part, to the January 2008 mark scheme being more complex and difficult to apply. The subjective experiences of the examiners are in line with the findings of quantitative analysis for PED4 (Section 1, above).

Table 11: The impact of online standardisation on second time examiners' quality of marking (percentage of examiners).

	Positive change	No change	Negative change	Total N
Awareness of the strengths and weaknesses in my marking	19.0	57.1	23.8	21
Depth of understanding of the mark scheme	17.4	43.5	39.1	23
Overall quality of my marking	14.3	57.1	28.6	21
Preparedness for marking	13.6	59.1	27.3	22
Internalising the mark scheme	4.8	76.2	19.0	21

As in June 2007, examiners commented that the lack of a face-to-face meeting to discuss the nuances of the mark scheme was detrimental to their quality of their marking:

"I feel really disadvantaged not having a standardisation meeting where I always left with a thorough understanding of the mark scheme." (PED4)

"I missed the face-to-face contact of an initial meeting and feel this negatively affected my preparedness to work." (PED4)

Theme 3: Support for Examiners

Table 12 shows how useful the examiners rated the support resources available to them during the second trial. In June 2007, the majority of examiners rated the online feedback on marking, the training day and user guide as useful or very useful (80.0%, 90.7% and 81.3% respectively). In January 2008, the online feedback (68.3%) and user guide (79.7%) were again amongst the

most highly rated support resources. The newly developed system preview component was also highly regarded, with 72.6% of examiners finding it either useful or very useful.

In June 2007, the most underutilised resources were the AQA email and telephone helpdesk and the online frequently asked questions. In January 2008, the vast majority of examiners (85.5%) did not attend the optional 'drop-in' training sessions. The AQA helpdesk email and the online frequently asked questions were again amongst the least used resources, 53.1 *per cent* and 47.6 *per cent*, respectively.

Examiner comments tended to focus upon difficulties experienced with the messaging service or helpdesk:

"Messaging service did not work properly. Queried to senior examiner took too long or didn't appear." (PED1)

"The messaging service didn't work EVER!" (3820/3)

"Helpdesk – 3 days to return emails and had to pester for a response." (GGB1)

"Helpdesk often unable to help and took time to get reply when time is allocated to mark and you can't because of technical fault it is frustrating." (PED1)

Table 12: Usefulness of online standardisation examiner support resources (percentage of examiners).

	Very useful	Useful	Not useful	Not used	Total N
Online feedback on marking	28.6	39.7	14.3	17.5	63
Query to Senior Examiner	19.0	20.7	15.5	44.8	58
Online standardisation user guide	12.5	67.2	10.9	9.4	64
AQA helpdesk telephone	9.4	35.9	17.2	37.5	64
System preview component	8.1	64.5	14.5	12.9	62
Messaging service	8.1	27.4	32.3	32.3	62
'Drop-in' training session	6.5	4.8	3.2	85.5	62
Senior Examiner training day	4.7	17.2	1.6	76.6	64
AQA helpdesk email	4.7	25.0	17.2	53.1	64
Online tutorial (video)	1.6	54.7	21.9	21.9	64
Online FAQs	1.6	38.1	12.7	47.6	63

Theme 4: Functionality of the System

Examiners were asked to rate their satisfaction with various aspects of the online standardisation system on a scale of 1 to 5 (1 = very dissatisfied, 5 = very satisfied). Between June 2007 and January 2008, the aspects of the system that examiners were most satisfied with varied substantially. In June 2007, the most highly rated functions of the system were the speed of feedback, the online messaging service, and the PEx's walk-through of the mark scheme. In January 2008, examiners were most satisfied with the screen-layout, the quality of graphics and the sizing of on-screen items (Table 13). The online messaging service received the lowest rating out of all the system features. These findings are not surprising given the development work on the software prior to the second trial (improved user interface with more space dedicated to scripts, new layout options to toggle between the script and mark scheme

and to display commentaries) and the technical difficulties experienced with the online messaging service.

A number of useful comments were made that should be used to inform the future development of the resource:

"On a number of occasions I could not change marks having put the wrong one in by mistake" (3820/3)

"There was an awful lot of waiting for information to be saved and loading, this made the process very tedious" (PED1)

"The fact the system has to register each mark individually really slows the process down" (PED1)

Furthermore, it appeared that whilst examiners were fairly satisfied with the speed of feedback (a mean satisfaction score of 3.41 and 3.21 in June 2007 and January 2008, respectively), they were discontent with when the feedback was received.

"I would have found it helpful if you could get feedback after every question within one paper marked, rather than having to wait for all the standardised batch to be finished before you got feedback and marks back" (3820/3)

"The fact we only get feedback at the end of the series of papers meant that it felt more like a test than a tool to improve our marking. Maybe it would be better to do it paper by paper" (PED1)

Table 13: Examiners' satisfaction with aspects of the online standardisation system (mean satisfaction score: 1=very dissatisfied, 5=very satisfied).

	Mean satisfaction score (sd)	**Total N**
Screen layout	3.47 (1.17)	60
Quality of graphics	3.42 (1.13)	59
Sizing of on-screen items	3.42 (1.20)	60
On-screen navigation	3.41 (1.19)	61
On-screen tools (e.g. being able to place marks anywhere)	3.27 (1.30)	59
Speed of feedback	3.21 (1.74)	58
Principal Examiner's walk-through the mark scheme	2.93 (1.68)	56
Online messaging	1.97 (1.71)	60

Perceptions of the system

Those examiners undergoing online standardisation for the first time were presented with a series of word dyads and asked to choose the word that best reflected their feelings towards online standardisation. Table 14 shows that examiners chose a selection of negative and positive words. Words such as 'convenient', 'interesting' and 'simple' were chosen alongside 'isolating', 'boring' and 'irritating'. It is somewhat revealing that 72.2% of respondents found online standardisation to be problematic. It seems likely that technical difficulties experienced with the system may have jaded their global impression of online standardisation.

Table 14: First time examiners perceptions of the system (percentage of examiners).

	Total N
Isolating (97.1%) - Inclusive (2.9%)	35
Irritating (82.4%) - Pleasing (17.6%)	34
Boring (77.8%) - Enjoyable (22.2%)	36
Convenient (73.7%) – Inconvenient (26.3%)	38
Problematic (72.2%) - Trouble-free (27.8%)	36
Uninspiring (67.6%) - Stimulating (32.4%)	37
Interesting (66.7%) - Dull (33.3%)	36
Patchy (62.9%) - Thorough (37.1%)	35
Simple (61.1%) - Complicated (38.9%)	36
Appropriate (60.0%) - Inappropriate (40.0%)	35
Suitable (59.4%) - Unsuitable (40.6%)	32
Inefficient (58.3%) - Efficient (41.7%)	36
Useful (58.3%) - Ineffective (41.7%)	36
Detailed (55.6%) - Superficial (44.4%)	36
Fussy (52.9%) - Streamlined (47.1%)	34
Informative (50.0%) - **Uninformative** (50.0%)	36

Examiners taking part in the trial for the second time were asked to compare their most recent experiences with those of the first trial by indicating whether they felt more, less or the same degree of confidence, frustration and so on. Nearly-two-fifths of examiners reported feeling more confident and less apprehensive using the system. However, 60.9 *per cent* of examiners reported feeling more frustration during the second trial. It seems likely that these frustrations stemmed from technical difficulties with the online resource e.g. the messaging service and the speed of the system.

Table 15: Second time examiners perceptions of the system (percentage of examiners).

	Less	Same	More	Total N
Frustration	8.7	30.4	60.9	23
Confidence	21.7	39.1	39.1	23
Segregation	8.7	52.2	39.1	23
Apprehension	39.1	43.5	17.4	23
Empowerment	17.4	65.2	17.4	23
Enthusiasm	26.1	69.6	4.3	23

DISCUSSION AND CONCLUSION

Analyses for the three conventionally marked components (3820/3, PED1 and PED4) revealed that the accuracy of matched examiners' first phase sample marking (in terms of percentage mean absolute mark difference) was greater for paper-based samples following face-to-face standardisation than online samples following online standardisation. For those examiners that

marked components in the first trial, the quality of their online first phase sample marking had remained stable.

The second phase sample analyses entailed two comparisons: 1) a comparison between paper-based samples in January 2007 (following conventional standardisation) and January 2008 (following online standardisation) and; 2) a comparison between paper-based and online samples following online standardisation in January 2008. On the whole, the outcomes for 3820/3 and PED1 were consistent with those of the first trial (Chamberlain, 2007). A non-significant difference was found between examiners' marking of the paper-based samples in the two January series, suggesting that mode of standardisation did not impact on the examiners' ability to mark paper-based second phase sample scripts. As to be expected from Chamberlain's first trial results and from the first phase sample results of the current trial, the accuracy of the examiners' second phase sample marking was greater for the paper-based than online sample following online standardisation. Should online second phase samples replace paper-based second phase samples in the future, consideration needs to be given to the criteria used to make examiner adjustments. It was reassuring that the online second phase sample marking of 3820/3 examiners' was comparable to that of June 2007.

The second phase sample outcomes for PED4 examiners differed to those for the other conventionally marked components. A significant difference was found between PED4 examiners' marking of paper-based second phase samples in January 2007 and January 2008. This finding is inconsistent with that of Chamberlain (2007) for PED4 in June 2007, and suggests that online standardisation may have impacted upon PED4 examiners' quality of paper-based second phase sample marking. In addition, the quality of the PED4 examiners online second phase sample marking was found to have declined since the previous trial. The PED4 examiners' marking of the online second phase sample remained less accurate than that of the paper-based second phase sample, following online standardisation.

It is possible that the PED4 examiners performed relatively poorly on the second phase sample, due to changes to the question paper and mark scheme and/or an enlarged workload. In January 2008 a number of PED4 examiners (7 out of 18) were also marking for PED1, which had been newly introduced to online standardisation. All PED4 examiners were also completing a third second phase sample, comprised of fifteen common paper-based scripts, as part of a research study. Efforts were made not to overload PED4 examiners (e.g. the 15 scripts included in the third second phase sample were deducted from the examiners' live marking allocations), but it is possible that these were insufficient. Examiners' that marked for both PED1 and PED4, for example, would have marked a total of five second phase samples over a relatively short period of time. These findings highlight the importance of not underestimating the impact of rolling-out online standardisation to additional components within the same specification, and having examiners complete research exercises during a live series. Having said this, it is important to note that the largest mean absolute mark difference of 3.55 *per cent* (Table 4) reported for the PED4 paper-based 2008 second phase sample (the basis for examiner adjustments), translates to only 2 out of 60 marks.

An exploration of examiner performance at total script level for the common online second phase sample scripts revealed that, on the whole, the examiners' marking was approximately in line with that of the PEx. This finding is reassuring, given the relatively large mean absolute mark differences observed at item level for the online second phase samples in all three conventionally marked components. Surprisingly, PED1 examiners' marking was most in line with the PEx at total script level (grand mean mark difference of 1.3%). PED1 was new to

online standardisation in January 2008; it may be worthwhile reviewing the PED1 online standardisation materials to identify pointers for best practice. The accuracy of the PED4 examiners' online second phase sample marking was found to have deteriorated at total script level (as well as item level) between trials. For the CMI+ components (CPT2, CPT5, GGB1), the limited data available from the Quality of Marking by Marker Reports suggested that mode of standardisation did not impact on the examiners' ability to mark seed items.

In many ways the comparisons discussed here are confounded, as they do not compare like-with-like. As outlined at the beginning of the paper, important procedural differences exist between paper-based and online sample marking. Under online standardisation, examiners mark a common set of scripts which have been previously marked by the PEx, and are devoid of any annotations. It seems almost inevitable that such conditions will lead to larger absolute mark differences. Analyses are currently underway to determine the extent to which the observed differences between the paper-based and online second phase sample scripts are a consequence of common versus partially self-selected scripts and marking onscreen versus on paper (Billington, In preparation). It is hoped that the findings of the analyses will inform the way in which examiners complete their second phase sample in the future and, thus, the basis for calculating examiner adjustments. Where comparisons have been made over-time, the use of different question papers and mark schemes pose a further threat to the internal validity of the analyses.

Since the January 2008 trial, a study has been conducted to explore marking reliability following conventional (face-to-face) and online standardisation for a GCSE History B component (Taylor, Chamberlain & Meadows, 2008). Eighty-nine examiners were randomly allocated to one of the two modes of standardisation. Prior to receiving standardisation all examiners marked a common set of cleaned scripts using the mark scheme only. These provided a baseline measure of the examiners' quality of marking. After undergoing either face-to-face or online standardisation examiners marked another set of common, cleaned scripts. The main finding of the study was that, after controlling for baseline marking reliability, there was no effect of mode of standardisation on examiners marking performance (Meadows & Taylor, 2008). This outcome should be given more credence than those for first and second phase sample data. The experimental design of the study meant that many confounding factors, such as common versus partially self-selected scripts, were controlled for. As such, the GCSE History study was a much more valid measure of the impact of online standardisation on examiners' marking reliability. Furthermore, analyses conducted by Taylor (2008) on enquiries after results data revealed that for those components standardised online in June 2007 and January 2008, there was no significant change in the number of mark changes following a re-mark request compared to the previous year. Such findings further support the introduction of online standardisation to a greater number of components.

The findings from the January 2008 evaluation questionnaire suggest that whilst online standardisation may be entirely feasible, many examiners have doubts regarding the transition from conventional (face-to-face) standardisation to online standardisation. They felt that the loss of face-to-face meetings, in which queries can be raised and clarification given, hindered the accuracy of their marking. In the January 2008 trial, technical difficulties experienced with the messaging service, speed of the system, and the helpdesk seem to have negatively impacted upon examiners global impressions of the online standardisation process. Such factors are rectifiable in the future as the online resource develops and improves.

Lucy Billington & Chloe Davenport
October 2008

REFERENCES

Baird, J. & Meadows, M. (Under review) *What is the right mark? Respecting other examiners' views in a community of practice.*

Billington, L. (In preparation). *Exploring second phase samples: What is the most appropriate basis for examiner adjustments?* AQA Research report.

Chamberlain, S. (2007). *E-standardisation pilot, summer 2007: Evaluation of first and second phase sample performance and examiner satisfaction.* AQA Research Committee Paper, RPA_07_SC_RP_061.

Cohen, J. (1992) A power primer. *Psychological Bulletin*, n112, pp.155-159.

Meadows, M & Taylor, R. (2008). *Online standardisation GCSE History Research Study – update report August 2008.* Internal AQA report.

Taylor, R., Chamberlain, S. & Meadows, M. (2008). *Comparing the effects of online and face-to-face training on marking reliability*. AQA Research Committee Paper.

Taylor, R. (2008) *The impact of online standardisation on enquiries after results*. AQA Research Committee Paper.