EXPLORING SECOND PHASE SAMPLES: WHAT IS THE MOST APPROPRIATE BASIS FOR EXAMINER ADJUSTMENTS?

Lucy Billington

ABSTRACT

In the UK, examination boards monitor the quality of their examiners' marking by sampling their work (scripts) at two different points during the marking period. The first sample is drawn almost immediately after the examiner has undergone training in the application of the mark scheme, whilst the second sample is taken approximately half way though the marking period. Both samples are over-marked by the examiner's Senior Examiner, but it is the second sample, known as the second phase sample, that is used in the examiner adjustments process. The first stage of the study reported here, explored 12 GCE examiners' marking reliability across 3 different types of second phase sample: 1) a conventional paper second phase sample 2) an online second phase sample and 3) a paper 'control' sample. Whilst the conventional paper second phase sample was self-selected by examiners and included their marks and comments, the latter two samples consisted of pre-selected, common scripts that were devoid of any marks and annotations.

The examiners' marking reliability for the conventional paper second phase sample differed from that for the online and paper 'control' second phase sample. These differences were largely attributed to the Senior Examiner being influenced by the marks and comments of the first examiner when reviewing the conventional paper second phase sample. Furthermore, the examiners comparable performance on the online and paper 'control' second phase sample added weight to the proposition that large absolute mark differences observed in previous research between conventional paper and online samples are attributable to fundamental differences in the way the two sample types are selected and presented rather than completing them online.

In the second phase of the study, the Principal Examiner re-marked a sub-sample of the live marking allocations for 5 of the examiners, with a view to exploring which of the 3 types of second phase sample was most representative of the examiners' overall marking reliability. It was found that the examiners' marking reliability for live scripts differed significantly from that for the conventional paper second phase sample. Current second phase sample procedures may underestimate the unreliability inherent in examiners' marking, and thus the adjustments needed to bring examiners' marking in line with the agreed standard.

INTRODUCTION

UK examination boards strive to ensure that examiners' marking is of a common high standard and free from bias. To this end, the Principal Examiner (PEx) trains examiners in the application of the mark scheme. In examinations with large entries, this examiner standardisation is achieved via a hierarchical system. The PEx trains Team Leaders (TLs) at a 'pre-standardisation' meeting. TLs then train and monitor the marking of small groups of examiners. TLs are responsible for ensuring that the standard set by the PEx is filtered down to examiners in their team.

This process, however, is not without difficulties and remedial measures are needed to detect and correct unreliable marking (Meadows and Billington, 2005). For examinations that are marked on paper rather than on-screen, two samples of each examiner's marking are evaluated to determine whether they have marked in accordance with the standard set by the PEx. The first phase sample (FPS) of 10 scripts is taken immediately after training to ensure that the standardisation has been successful. A second phase sample (SPS) is taken approximately half way through the marking period and consists of 50 scripts, selected by the examiner. In the first instance, the TL over-marks 15 of the scripts. If the marking seems problematic, i.e. the total script marks are outside an agreed tolerance, an additional 10 scripts are over-marked. This sample of 25 re-marked scripts forms the basis on which decisions are made about examiner adjustments. Examiners deemed to have been consistently lenient or severe in their marking have an adjustment applied to the scripts in their marking allocation. The quality control processes relating to examinations marked on-screen are quite different and involve 'seeded' items since marking occurs at item level.

Recently, AQA has begun to standardise examiners remotely via online systems rather than through face-to-face training. This has opened up the possibility of different methods of collecting SPS information and of calculating examiner adjustments. In trials of online standardisation for components that were marked on paper, each examiner submitted a conventional paper SPS on which examiner adjustment decisions were made and an online SPS. The online SPS had been pre-selected and assigned a mark by the PEx. Examiners could not see the marks and annotations of the PEx. Figure 1 summarises the main distinctions between paper-based and online samples. Analyses revealed that the accuracy of examiners' marking appeared significantly greater when paper rather than online SPS samples were completed (Chamberlain, 2007, Billington, 2008). However, the number and kind of components included in the trials was limited and examiners were more familiar with conventional SPS procedures.

Figure 1: A summary of the procedural differences between paper-based and online samples

Paper	Online
• Self-selected by the examiner from their marking allocation • Re-marked by TLs on paper • Includes marks and annotations of the first examiner (the Assistant Examiner)	• Pre-selected and assigned a 'true' score by the PEx • Re-marked by examiners onscreen • Excludes mark and annotations of the first examiner (the PEx)

Research suggests that it is likely that in the paper SPS TLs' re-marking was influenced by the marks and annotations of the examiners, giving the appearance of greater marking accuracy. Indeed, Murphy (1979) compared the reliability of examiners' marking when scripts had the first examiners' marks and comments on them and when scripts had been 'cleaned'. Removing the marks and comments of the first examiner approximately doubled the absolute mark differences observed between the mark awarded by the first examiner and those awarded by the re-marking examiner. More recently, Baird and Meadows (under review) reported greater discrepancy between examiners' and Senior Examiners' marking of photocopied 'clean' scripts than 'live' annotated scripts. Since there was no evidence that the photocopied scripts were more difficult to mark than the live scripts, they suggested that the Senior Examiners were probably being influenced by the marks and comments of the original examiner when reviewing

their marking. They also found evidence that the examining personal responsible for assigning the 'true' mark to scripts included in samples may have introduced variability. Examiners' marking often matches more closely that of their TLs than that of the PEx suggesting a failure of pre-standardisation. Thus, it is unsurprising that previous research has observed greater agreement between the marks awarded by the examiner and their TL (paper samples) than between those awarded by the examiner and the PEx (online samples).

Different SPS procedures are likely to suggest different levels of examiner accuracy, but it is not clear which type of SPS is most representative of an examiner's live marking. SPSs are the primary evidence used in deciding the outcome of the examiner adjustments procedure. Since any adjustment is applied to all candidates (within a specified range), it is imperative that the SPS represents the quality of the examiner's marking across their entire marking allocation. Otherwise, adjustment decisions may be misguided and place some candidates at an unfair advantage and others at an unfair disadvantage.

Furthermore, it is apparent that procedural differences identified between paper-based and online samples confounded previous analyses; it is unknown whether the larger absolute mark differences observed for online samples are attributable to differences in sample selection (self-selected versus pre-selected) and script presentation (annotated versus clean scripts) or whether there was some effect of marking scripts online rather than paper. It would be valuable to consider a paper 'control' sample, consisting of common scripts, pre-selected and assigned a 'true' mark by the PEx, hence, disentangling the effect of script selection and script presentation from the effect of marking on paper or online.

This study sets out to explore the form of SPS which is most closely related to examiners' live marking, and thus, the most appropriate basis for the calculation of examiner adjustments.

METHOD

GCE Sport and Physical Education (PED4) formed the focus of the enquiry. Examiners were standardised online and had been in one previous series. It was hoped that some familiarity with the online standardisation system and completing an online SPS would facilitate a fairer assessment of 1) the type of SPS that best represents examiners' live marking, and 2) the impact of completing the SPS online rather than on paper on examiners' quality of marking, following online standardisation.

In the first stage of the study, PED4 examiners underwent online standardisation and completed 3 SPSs. Examiners completed an online SPS and a conventional paper SPS. The latter was used in the examiner adjustment process. The third SPS was a paper 'control' sample. This sample consisted of 15 scripts, pre-selected and allocated a mark by the PEx. The PEx was asked to use the same selection criteria for the paper 'control' SPS as for the online SPS. The purpose of the paper 'control' SPS was to separate the effects of script selection (self-selected versus pre-selected) and script presentation (annotated versus clean scripts) from the effect of sample completion method (online versus paper). Stage 1 was, thus, a within-subjects design and consisted of three conditions.

Figure 2: A summary of the second phase samples completed by PED4 examiners in February 2008

Condition	Script Selection	Script Presentation	Sample Completion Method
1. Conventional paper	Self-selected	Annotated	Paper
2. Online	Pre-selected	Clean	Online
3. Paper 'control'	Pre-selected	Clean	Paper

Initially, the design of the study was intended to be counterbalanced, with examiners completing the 3 SPSs in different orders e.g. 123, 132, 231, 213, 312, 321. Unfortunately, operational constraints, such as the deadline for examiner adjustments and the 2 day window available for examiners to access the online standardisation system and complete the online SPS, meant that it was not possible to control for order effects. It is almost certain that the examiners completed the conventional paper SPS first, followed by the online SPS, and finally the paper 'control' SPS. Thus, carry-over effects were also of concern. It was thought that the examiners' marking might improve with each SPS marked, as they became more familiar with the question paper and mark scheme. Whilst such factors may have jeopardised the internal validity of the study, it should be noted that it was imperative that the study did not interfere with AQA's operational activities and, ultimately, the timely delivery of candidates' examination results.

In the second stage of the study, the PEx was required to re-mark a sub-sample of some of the PED4 examiner's live marking allocations. The purpose of the re-marking exercise was to assess which of the 3 SPSs was most representative of the examiners' live marking. On the basis of stage 1 data analyses, 5 examiners were selected for inclusion in stage 2. These 5 examiners showed variability in the accuracy and consistency of their marking across the 3 SPSs. In total, the PEx re-marked 157 scripts. This number of scripts approximated to the average marking allocation for PED4 in the February 2008 examination series, and equated to one quarter of each examiner's live marking allocation (see Table 2). The scripts were randomly selected from a list of each examiners live marking allocation. To prevent the PEx from being biased by the marks and comments of the first examiner the scripts were 'cleaned' prior to re-marking.

Restricted access to candidates' scripts (due to enquiries after results and access to scripts deadlines) and the PEx's commitments for the June 2008 examination series prevented the re-marking exercise from taking place until August 2008. The PEx may have forgotten certain details of the mark scheme during the lapse of time between the examination and the re-marking exercise, potentially giving rise to inaccurate results. For this reason, the PEx was asked to re-familiarise himself with the February 2008 mark scheme and question paper before starting to re-mark any scripts.

RESULTS

Stage 1

In total, 12 PED4 examiners fully completed all 3 SPSs in February 2008. The quality of the examiners SPS marking was explored by comparing each examiner's marking with that of the Senior Examiner. It should be noted that for the conventional paper SPS the examiners' marks have been compared with those of their TL, where as for the online SPS and paper 'control' SPS their marks have been compared with those of the PEx. TLs and the PEx are, where appropriate, collectively referred to as 'Senior Examiners' in this paper.

Specifically, two measures of marking reliability were explored:

1. The absolute mark difference (AMD) between an examiner's mark and the Senior Examiner's mark.
2. The correlation between the marks awarded by an examiner and the Senior Examiner.

Absolute mark discrepancies provide a better measure of marking reliability than raw mark discrepancies, as with the latter positive differences are allowed to counteract negative differences. Whilst absolute mark discrepancies between an examiner and the Senior Examiner measure marking *accuracy*, the correlation between the marks awarded by an examiner and the Senior Examiner measures marking *consistency*. Arguably, in examination systems where examiners mark whole papers rather than individual scripts, the latter is more important. A consistent but severe or lenient examiner can have his/her marks adjusted (Baird & Mac, 1999).

For each SPS type and individual examiner, Table 1 shows the total AMD (across 15 scripts) and the correlation between the examiner's marks and those of the Senior Examiner. The AMD is reported as a percentage of the absolute mark difference possible eg if the Senior Examiner gave a mark of 25, with a maximum mark of 64, an examiner could be up to 39 marks away from the Senior Examiner. All examiners marked the conventional paper SPS with the greatest degree of accuracy (as illustrated by the low percentage AMDs); with 11 out of 12 examiners marking this sample most consistency with the Senior Examiner (as illustrated by the high correlations). Cohen and Holliday (1982) suggest that 0.19 and below is a very low correlation; 0.20 to 0.39 is low; 0.40 to 0.69 is modest; 0.70 to 0.89 is high; and 0.90 to 1 is very high. Thus, all correlations were either high or very high. The smallest correlation was observed for an examiner's paper 'control' SPS (examiner 8, 0.77), whilst perfect correlations were observed for 2 examiners' conventional paper SPS (examiners 9 and 12). For 8 out of 12 examiners the percentage AMD increased from that on the conventional paper SPS when using online SPS and increased further using the paper 'control' SPS. The same pattern was observed for 10 out of 12 examiners with regard to the correlation with the Senior Examiner.

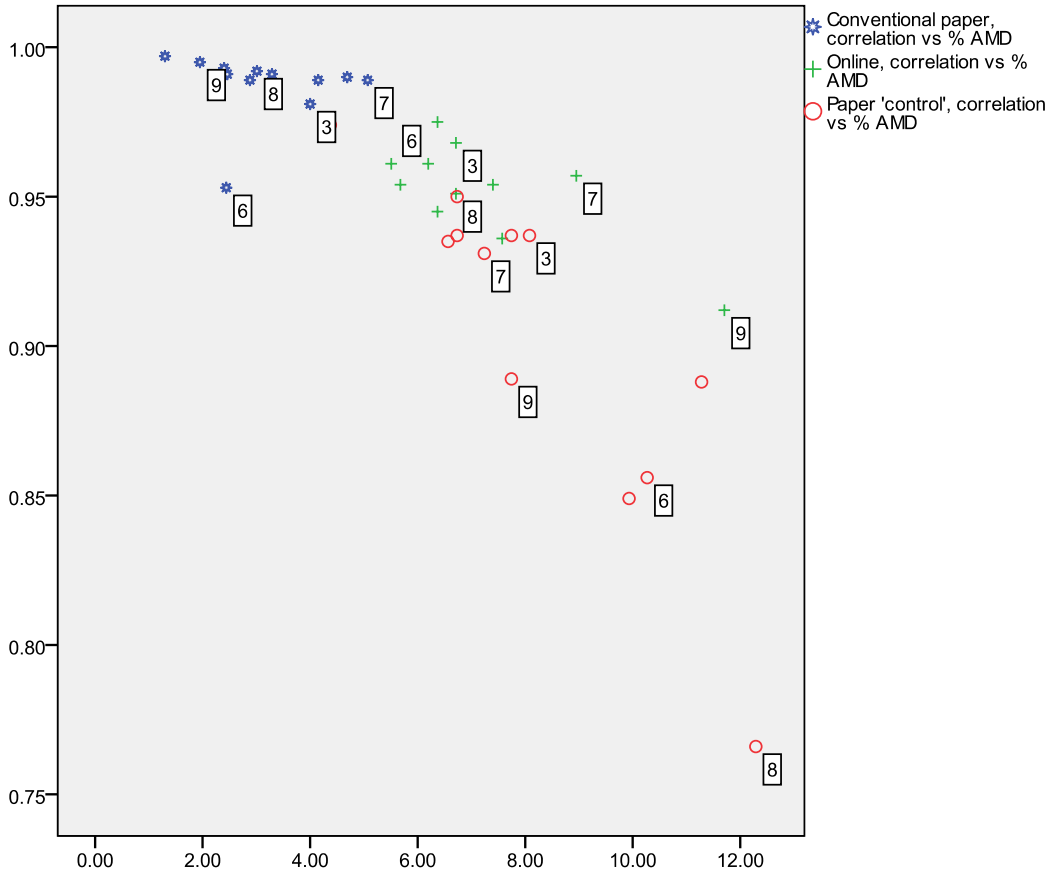
Table 1: Percentage AMD and correlation by second phase sample type

	Conventional paper		Online		Paper 'control'	
Examiner ID	**% AMD**	**Correlation**	**% AMD[1]**	**Correlation**	**% AMD**	**Correlation**
12	1.30	1.00	7.40	0.95	6.73	0.94
9	1.95	1.00	11.70	0.91	7.74	0.89
4	2.40	0.99	5.68	0.95	4.38	0.97
6	2.44	0.95	6.20	0.96	10.27	0.86
11	2.46	0.99	5.34	0.98	7.74	0.94
2	2.88	0.99	6.37	0.95	6.57	0.94
8	3.01	0.99	6.71	0.95	12.29	0.77
5	3.29	0.99	6.37	0.98	6.73	0.95
3	4.00	0.98	6.71	0.97	8.08	0.94
1	4.15	0.99	5.51	0.96	11.28	0.89
10	4.69	0.99	7.57	0.94	9.93	0.85
7	5.07	0.99	8.95	0.96	7.24	0.93

[1] The online second phase sample did not include marks for Quality of Written Communication (QWC). The maximum mark for the paper was 64, with 4 marks being awarded for QWC. Consequently, a maximum mark of 60 was used when calculating the AMD as a percentage of the maximum mark difference possible for the online second phase sample.

For each SPS, Figure 3 plots each examiner's percentage AMD against the correlation of their marks with those of the Senior Examiner. This further highlights the tendency for the examiners to appear to mark the conventional paper SPS most reliably and the paper 'control' second phase sample least reliably.

Figure 3: Examiner's percentage AMD and correlation with the Senior Examiner for each second phase sample type



Repeated measures ANOVAs were conducted to compare the examiner's marking performance across the 3 SPS types[2]. A significant difference between the percentage AMDs achieved by the examiners for the 3 SPS types was found (F (2, 22) =28.02, p<0.001). Moreover, 72 *per cent* of the variation in the examiner's AMD could be explained statistically by the type of SPS (eta-squared=0.718). Pairwise comparisons for the main effect of SPS type, corrected using a Bonferroni adjustment, revealed that the significant main effect reflected a significant difference between the conventional paper SPS and the online SPS (p<0.001) and the conventional paper SPS and paper 'control' SPS (p<0.001). Interestingly, a significant difference was not found between the examiners' mean percentage AMD for the online SPS and paper 'control' SPS (p=0.565). In other words, the examiners' apparent marking accuracy differed between the conventional paper and online SPS and the conventional paper and paper 'control' SPS, but was comparable for the paper 'control' and online SPS.

[2] A Fisher transformation was applied to all correlation coefficients to allow their use in ANOVA analyses.

The correlation of the examiner's marks with those of the Senior Examiner was also found to differ significantly by SPS type (F (2, 22) =49.54, p<0.001). Eighty-two *per cent* of the variation in the examiner's correlation with the Senior Examiner could be explained statistically by the type of SPS (eta-squared=0.818). Pairwise comparisons revealed that all three mean correlations were significantly different from each other. Although all examiners exhibited high to very high correlations, it seems that their marking varied in consistency across the 3 SPSs.

Stage 2

Stage 2 of the study consisted of the PEx re-marking a sub-sample of some examiner's live marking allocations. The purpose of the re-marking exercise was to explore which of the 3 SPSs most closely resembled the examiners' marking reliability for live scripts. Five out of the 12 examiners were selected to be included in the re-marking exercise. The 5 examiners are labelled by Examiner ID in Figure 3 above. Each of the examiners exhibited considerable variability in their marking of the 3 SPSs. This was important as, if the quality of an examiner's marking had been too similar for each of the SPSs, it would have been difficult to determine which of the SPSs was most representative of their live marking. In total, the PEx marked 157 scripts – one quarter of each examiner's live marking allocation (Table 2).

Table 2: Total marking allocation for each examiner and the number of scripts re-marked by the PEx

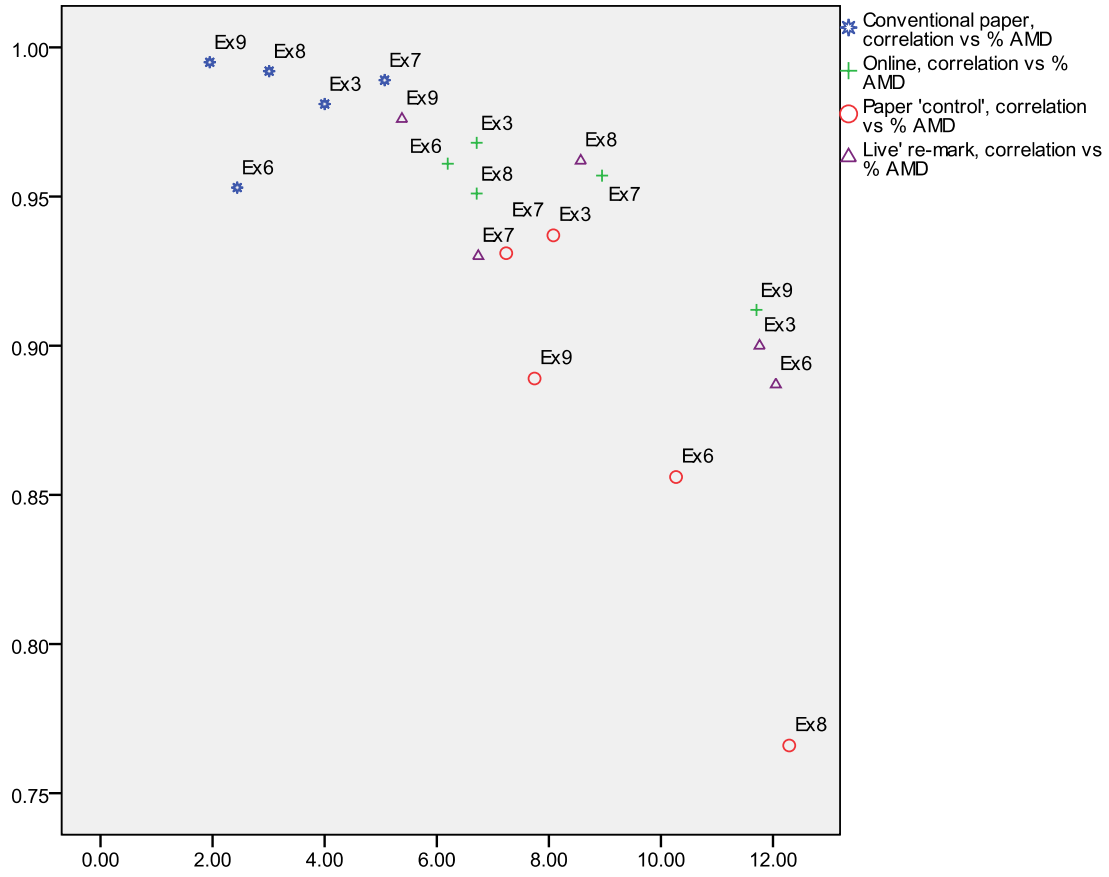
Examiner ID	Total marking allocation	No of scripts re-marked
3	132	33
6	80	20
7	137	34
8	141	35
9	140	35
Total	**630**	**157**

Table 3 shows each examiner's AMD as a percentage of the maximum mark difference possible and the correlation of their marks with those of the Senior Examiner for the scripts included in the re-marking exercise and for the 3 SPSs. These data are also plotted in Figure 4. The small sample size means that caution must be exercised when interpreting the data. Worryingly, 2 examiners (Examiners 3 and 6) marking of live scripts appeared less accurate than any of their marking of the 3 SPSs. Likewise, an examiner (Examiner 3) appeared to mark their live scripts less consistently. Figure 4 shows that measures of the examiners' marking reliability clearly deteriorated between SPSs (conventional paper being most reliably marked and paper 'control' least reliably marked), but these do not have a strong and consistent relationship with the reliability of their live marking.

Table 3: Percentage AMD and correlation for the live re-marked scripts and each second phase sample

	Live re-mark		Conventional paper		Online		Paper 'control'	
Examiner ID	% AMD	Correlation	% AMD	Correlation	% AMD	Correlation	% AMD	Correlation
9	5.38	0.98	1.95	1.00	11.70	0.91	7.74	0.89
7	6.74	0.93	5.07	0.99	8.95	0.96	7.24	0.93
8	8.57	0.96	3.01	0.99	6.71	0.95	12.29	0.77
3	11.76	0.90	4.00	0.98	6.71	0.97	8.08	0.94
6	12.05	0.89	2.44	0.95	6.20	0.96	10.27	0.86

Figure 4: Examiner's percentage AMD and correlation with the Senior Examiner for the live re-marked scripts and each second phase sample type



Repeated measures ANOVAs revealed a significant main effect of sample type (live re-marked scripts and the 3 SPS) on the examiners' marking accuracy (F (3, 12) =5.84, p=0.011, eta-squared=0.593) and marking consistency (F (3, 12) =9.49, p=0.002, eta-squared=0.704). More importantly, simple contrasts revealed that only the conventional paper percentage AMD and conventional paper correlation were significantly different from those observed for the live re-marked scripts. In other words, the online and paper control SPSs seemed to better estimate the reliability of live marking than the conventional paper SPS.

DISCUSSION & CONCLUSION

As would have been predicted from previous research on the effects of visible marks and annotations on re-marking examiners' behaviour, examiners' marking of conventional paper SPSs appeared more accurate and more consistent than that of online and paper control SPSs. Further, examiners' marking accuracy for the online and paper 'control' SPSs was comparable. This suggests that the larger absolute mark differences observed for online than conventional paper SPS in previous research (Chamberlain, 2007; Billington, 2008), were mainly a consequence of the scripts being pre-selected by the PEx and devoid of any marks and comments, rather than an effect of online marking.

However, it is also likely that motivational factors influenced the examiners' quality of marking across the 3 SPSs. Examiners' performance is judged on their marking of the conventional paper SPS, which also informs examiner adjustment decisions. One would, thus, have expected examiners to have taken considerable care when selecting their scripts for inclusion in

this SPS. Examiners knew that the paper 'control' SPS was part of a research study, and it seems possible that they were less conscientious in their marking of this sample. An alternative explanation concerns changes in marking consistency over-time. Pinot de Moira, Massey, Baird and Morrissy (2001) found a small, but statistically significant, change in the accuracy of marking over the marking period, with examiners' marking tending to increase in severity. Examiners almost certainly completed the conventional paper SPS first, followed by the online SPS and, lastly, the paper 'control' SPS, and this may have contributed to the declining marking accuracy.

The findings from the re-marking exercise suggest that conventional paper SPSs underestimate the unreliability inherent in examiners' marking. The examiners' marking accuracy and marking consistency for their live scripts did not significantly differ from that displayed for the online and paper 'control' SPSs, but did significantly differ for the conventional paper SPS scripts. However, a number of limitations must be borne in mind. Firstly, the small number of examiners involved in the re-marking exercise makes it difficult to interpret and generalise the findings. Secondly, it is possible that the outcomes were a product of the study design. Examiners' were selected for inclusion in the re-marking exercise, because their marking was particularly erratic across the 3 SPSs. Had a different 'type' of examiner been included in this stage of the study, the findings may well have been different. Moreover, the PEx re-marked scripts in a cleaned state. Since removing marks and comments leads to greater discrepancies between the marks awarded by the first examiner and re-marking examiner, it is not surprising that the conventional paper SPS sample appeared to underestimate the examiners' marking unreliability.

A fundamental question that needs to be addressed concerns the most appropriate way of measuring examiners 'real' marking reliability. Had the PEx re-marked annotated scripts it is probable that the conventional paper SPS would **not** have appeared to have underestimated unreliability. Indeed, there is some doubt in the literature over whether re-marking clean scripts represents the most optimal check on examiners' marking. Community of practice literature would suggest that taking another examiner's marks into account is an entirely legitimate process of reaching an agreement about examination standards (Meadows & Baird, under review).

Should SPSs consisting of pre-selected, common scripts be used in the future, it is arguable that they would facilitate a more rigorous and fairer strategy for monitoring examiners' marking, and thus, applying adjustments. They provide a common basis from which to compare an examiner's marking for the same paper. Furthermore, pre-selected common scripts have the advantage that Senior Examiners spend less-time re-marking the scripts of their colleagues. Meadows (2006), however, noted a number of disadvantages of using pre-selected, common scripts in monitoring examiners' marking. Firstly, it takes time to select and prepare these scripts, lengthening pre-standardisation meetings. Secondly, resources are spent on duplicate marking, rather than on examiners' marking of 'live' scripts.

In order to determine the type of SPS that most accurately predicts an examiner's marking of live scripts in their allocation, a similar, but much larger scale study would be required. This study has, however, raised some doubts over whether the conventional paper SPS is the most appropriate basis on which to base examiner adjustments. Electronic marking systems monitor examiners' marking in 'real' time and no adjustments are required. In the long term, electronic marking offers a more sophisticated solution to the dilemma of how best to monitor the marking of examiners for national examinations.

Lucy Billington
June 2009

REFERENCES

Baird, J. & Mac, Q. (1999) How should examiner adjustments be calculated? A discussion paper. Internal paper, AEB. RC13

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

Billington, L. (2008) *Online standardisation trial, winter 2008: Evaluation of examiner performance and examiner satisfaction*. Internal paper, Assessment & Qualifications Alliance. RPA_08_LB_RP_054.

Chamberlain, S. (2007). *E-standardisation pilot, summer 2007: Evaluation of first and second phase sample performance and examiner satisfaction.* Internal paper, Assessment & Qualifications Alliance. RPA_07_SC_RP_061.

Cohen, L. & Holliday, M. (1982) *Statistics for Social Scientists*. London: Harper & Row.

Meadows, M & Billington, L. (2005) *A review of literature on marking reliability*. Internal paper, Assessment & Qualifications Alliance. RPA_05_MM_RP_05.

Meadows, M. (2006) *The use of 'live' versus photocopied scripts in the first sample of marking standardisation.* Internal paper, Assessment & Qualifications Alliance. RPA_06_MM_RP_019a.

Murphy, R.J.L. (1979) Removing the marks from examination scripts before re-marking them: Does it make any difference? *The British Journal of Educational Psychology*, v49 pp.73-78.

Pinot de Moira, A., Massey, C., Baird, J. & Morrissy, M. (2001) *Marking consistency over time*. Internal paper, Assessment & Qualifications Alliance. RC129.