



DOES MARKING IMAGES OF ESSAYS ON SCREEN RETAIN MARKER CONFIDENCE AND RELIABILITY?

ABSTRACT

Encouraged by the finding, from mark/re-mark data, of impressive marking reliability in the major assessment in English for 16 year-olds in the UK, a sample of students' scripts was transferred from paper to scanned images and presented for electronic marking. The awarding body responsible for these assessments in English (AQA) uses electronic marking for short and structured questions in many examinations but not for English, where the scripts include both short answers and essays. Marking essays and other long answers on screen is seen as challenging, particularly in English where an holistic view is required to determine an appropriate level and mark. The marks awarded electronically, using an enhancement to the e-marking software that imitates paper-based marking by giving markers a facility to annotate script images with marks and comments, were analysed using the same variety of approaches to marking reliability as used in an earlier paper-based investigation, with results for the essay question that were rather poorer than on paper. The apparent interference with the markers' normal essay marking is a reminder that developments in technology for assessment purposes may introduce gains in efficiency but detract from reliability and validity, although a designed study such as this cannot capture all the benefits that may be gained from e-marking, particularly real time monitoring of examiners and early intervention should their marking deviate from the mark scheme.

INTRODUCTION

Electronic marking of scanned images of responses is being introduced rapidly to UK GCE and GCSE examinations. In one awarding body, the Assessment and Qualifications Alliance (AQA), over 130 components were marked electronically in the 2008 examinations, using the CMI+ application in the DRS e-Marker® suite. The annotation tool is a CMI+ development designed to draw longer answer responses into electronic marking. It allows examiners to place marks and include comments on the screen image at appropriate points in the candidates' responses. Examiners annotate scripts in normal paper-based marking, and it proves useful if not essential for post-results enquiries and appeals that examiners leave such a trail.

Crisp and Johnson (2007) speculate on the function served by the annotation process during the marking process. They see it as serving two purposes: to help explain examiners' decisions (useful for post-results queries), and to support their judgements and decision-making. Their study analysed a small number of scripts across two different subjects to look for any patterns in the use of annotations and to survey which annotations were subject-specific and which were more general. The findings showed that the frequency and type of annotations used differed between the subjects, and that examiners within a subject team drew from a

common pool of annotations. They concluded that understanding the purpose of making annotations in individual subjects might help inform the preparation of examiners and be of use to future developments in marking. They cite evidence from Bramley and Pollitt (1996), who carried out an experimental study on the use of annotations in Key Stage 3 English marking, that it is important for examiners' perceptions of their own marking that they are able to annotate scripts, feeling that it improves their marking and reduces the subjectivity of their judgements. The evidence from the marking, however, was that there were no significant differences in marking accuracy, although using some annotation rather than none tended to make marking a little more consistent. Raikes, Greatorex and Shaw (2004) also note that examiners report feeling that they need to be able to annotate scripts as they e-mark *'in order to mark properly'*. Johnson and Shaw (2008) highlight the potential role of annotating on reading comprehension processes, citing literature about linguistics and annotating practices to suggest a link between them. Annotation thus appears to be a feature of conventional marking that serves an essential function. It may involve no more than ticks, although in a pilot of on-screen marking of pieces of extended writing in diagnostic English tests for 14 year-olds, Shaw (2008) reports that examiners preferred a condition of 'sophisticated' to 'simplified' annotation. The software used in the pilot imposes a standardised set of electronic annotations; examiners using the simplified suite found them restrictive. However no statistically significant mark differences were found in four marking conditions that consisted of either sophisticated or simplified annotation combined with either paper-based or on-screen marking.

Fowles and Adams (2005) considered differences in the assessment process when e-marking replaces paper-based marking, and commented on the perceived need to annotate scripts. They observed that CMI+ from its first usage by AQA had allowed examiners a limited facility to add comments, which had removed a little of the difference between the two forms of marking, but that there was every likelihood that further technical developments would allow examiners *'to make whatever annotations and comments that they wish, according to their current practices'*. (p15). CMI+ with the addition of an annotation tool was duly introduced in 2006 for a few GCSE components. Annotations of any kind are permitted, and the annotated script image is stored together with the examiner's marks, which are all recorded wherever on the screen image the examiner chooses to place them, and absorbed into the item and component totals.

AQA has been cautious in its introduction of e-marking, and has required evidence that the quality of marking is preserved in the transition from paper to marking images on screen. None of the components initially introduced to e-marking with the annotation facility required lengthy written responses from candidates and therefore a trial of e-marking of longer, essay style responses was suggested in order to demonstrate more fully the acceptability, or otherwise, of the annotation feature. A GCSE English paper that had been included in an earlier investigation of the reliability of marking in the current AQA GCSE English examinations (Fowles, 2006) was considered suitable for such a trial.

English is just one subject where the nature of candidates' responses has been regarded within AQA as challenging for e-marking. Marking essays in English has long been seen as a relatively unreliable activity, as an early paper on the GCE O-level examination attests: 'one cannot expect high correlations between two examiners marking the same set of compositions' (Hewitt, 1967, p5). (Measures taken to promote high levels of marking reliability can prove controversial however, as recent comment provoked by AQA's Chief Examiner for GCSE English in relation to the mark scheme demonstrates (Times Online, 2008).) The earlier (paper-based) investigation (Fowles, 2006) compared AQA's two GCSE English syllabuses, which differ with respect to their approaches to differentiation in the written papers (for Specification A

differentiation is by outcome, with a high proportion of similar but not identical questions for Foundation and Higher tier candidates, while in Specification B there are separate tasks for the two tiers). The script samples were 'cleaned' for marking in a re-marking exercise involving the Principal Examiners, whose marks were adopted as the best or 'true' marks, for comparison with the actual or 'live' marks awarded in the examination. Meadows & Bair (in press) discuss how the marks of such senior examiners are not the 'true' marks of classical test theory, which would result from the pooled judgement of an infinite number of markers (Spearman 1904a, 1904b, 1927) in what may be seen as a consensus model of the true mark; instead they denote a hierarchical definition of true mark, reflecting the operational reality of examinations in the UK, where the mark a candidate receives is likely to be the result of the judgment of one examiner or perhaps two. A number of approaches to investigating mark/re-mark reliability on paper was used, and collectively led to the conclusion that marking reliability was strongest in the Higher tier of Specification A, where for Paper 1 there was an impressive correlation of 0.95 between the 'live' and 'true' marks. Paper 1 was therefore selected for the present investigation of electronic marking, which had the two aims of (a) exploring whether the enhancement of CMI+ with an annotation facility could support marking in components where the nature of candidates' responses is regarded within AQA as challenging for e-marking and (b) comparing measures of paper-based and electronic marking reliability in GCSE English.

METHOD

In order to compare marking reliability for paper-based and e-marking using CMI+ with the additional annotation feature, the investigation had a between-subjects design (where the 'subjects' were GCSE English scripts) in which the independent variable was the medium, paper or electronic, for marking two random samples of scripts. The dependent variable was the reliability of the marking in each medium of a panel of six examiners, one of whom was the Principal Examiner. Marking reliability was gauged by various measures of the relationship between the Principal Examiner's ('true') marks and those awarded by the other examiners under the two marking conditions.

The scripts used in the investigation were drawn from the June 2005 GCSE English Specification A Higher tier examination, which as noted above had the strongest measures of marking reliability in the earlier investigation comparing paper-based marking of AQA's two GCSE English specifications (Fowles, 2006). Paper 1 was selected for the exercise; it is made up of five part questions in Section A, with mark allocations ranging from 3 to 8 marks and a total of 27 marks, followed by a choice of one from four essay questions in Section B. The essay is also allocated 27 marks, of which 9 are for spelling, punctuation and grammar.

Samples of scripts were first 'cleaned' using an electronic scanner and colour filter to remove the original examiners' marks written in red ink. A first randomly selected set of 45 scripts was printed and prepared for electronic marking using CMI+. The CMI+ application requires candidates' responses to be in predetermined locations, as in a combined question paper and answer booklet. The preparation of the scripts for electronic marking therefore involved extensive cutting and pasting into such a booklet. Each question/part question was allocated as much space as was needed to accommodate the longest written response in the set of 45 cleaned scripts. Four pages out of fourteen for the whole paper were for the essay question in Section B, although very few candidates wrote at this length, a page and a half of writing being typical. The prepared scripts were then scanned. Inspection of the script images on screen cast some doubt on the legibility of two scripts that had not been written in a suitable ink, and the prepared copies were therefore withdrawn in order that the extra variable of poor quality script evidence should not be introduced into the comparisons of marks awarded. The first

three scripts were identified as training and re-familiarisation material, which left 40 scripts to contribute data to the investigation. A second randomly selected set of 35 Paper 1 scripts was also prepared for re-marking on paper in the normal way.

The Principal Examiner, the Chief Examiner, a Team Leader and three assistant examiners carried out the marking after being introduced to the CMI⁺ operations and spending some time in re-familiarisation with the paper, source material and mark scheme. All six examiners were aware that they could mark the questions in whatever order they preferred. They understood that the first three responses for each question were included for the purpose of practice and re-familiarisation. They each marked all 78 scripts, including the three used for CMI⁺ training and re-familiarisation. The electronic marking was of necessity completed in the AQA office, and marking of some scripts from the paper-based set was also completed in the same office conditions, mostly interspersed with the electronic marking by way of a break from the screen. The paper-based marking was then completed at home.

The examiners were separately interviewed on completion of their electronic marking and invited to give their individual impressions and comments on the experience. The interview included views on the electronic marking of each of Sections A and B, and on segmented (question by question) marking with standard quotas as in CMI⁺. They were also asked about their attitude to, and use of, the annotation facility.

Each examiner's marks on the two sets of sample scripts are available for comparison with the Principal Examiner's 'true' marks in each medium, i.e. in both the paper and the electronic contexts. The assumption is made, as it had been in the earlier investigation, that the Principal Examiner's marks can be regarded as the 'true' mark. The earlier paper compared the marks awarded to each set of scripts by each examiner, including the absolute mark differences and the coefficients of correlation between the 'true' marks and the individual examiner's marks, at item, section and total mark level as appropriate.

RESULTS

The mean and standard deviation of the average absolute mark differences for each section and for the whole paper are given in Table 1, which includes the 95% confidence intervals ranges. The transfer to CMI⁺ from paper-based marking did not produce statistically significant changes (t-tests, with d.f.=73: for Section A, $t=1.51$; Section B, $t=-1.71$; total: $t=1.32$).

Table 1 Mean and standard deviation of absolute mark differences by medium

(a) Paper-based marking (N=35)

	mark allocation	mean AMD	s.d.	95% CI range	
Section A	27	2.82	1.20	2.40	3.23
Section B	27	2.27	1.03	1.92	2.63
Total	54	4.33	2.01	3.64	5.02

(b) Electronic marking with CMI⁺ (N=40)

	mark allocation	mean AMD	s.d.	95% CI range	
Section A	27	2.47	0.79	2.21	2.72
Section B	27	2.77	1.40	2.32	3.21
Total	54	3.78	1.55	3.28	4.28

The mean Paper 1 absolute mark difference in the earlier paper-based study (Fowles, 2006) was only 1.66, or just 3 per cent of the mark allocation. While the outcome of the present exercise was disappointing in comparison with the earlier study, it is nevertheless encouraging that the mean absolute mark difference, at 7 or 8 per cent, is still lower than might be popularly expected, because of supposed subjectivity, for marking in English, especially allowing for the time that had passed since the examination had been 'live' marked.

Although the paired comparison tests by examiner of the 'true' and individual examiners' mean marks, ignoring the direction of the mark differences, found some statistically significant differences between the paper-based and the electronic marks, there was no consistent pattern, whether at the level of the individual questions, section totals or overall total. A pattern was however detected in the coefficients of correlation between the Principal Examiner's 'true' marks and the individual examiner's marks. These are given in Table 2 by section and for the overall total, for marking by each examiner (a) on paper and (b) with CMI⁺. (Figures for the two senior examiners are in columns labelled SE1 and SE2 and those for the three assistant examiners are in columns labelled AE1 to AE3.) The coefficients for the whole paper marked on paper are all in the range 0.72 to 0.82 (lower than found previously (Fowles, 2006)¹) but are lower for each examiner for electronic marking, in the range 0.59 to 0.80. The Section A coefficients are similar for marking on paper and with CMI⁺, but the Section B coefficients are considerably lower for all five examiners with CMI⁺, being in the range of 0.29 to 0.43 compared to 0.60 to 0.78 for marking on paper.

Average correlations have been calculated over the five examiners, together with confidence intervals at the 95 per cent level, using the Fisher's Z transformation. The mean and the upper and lower limits are given in the final columns in Table 2. The ranges of values within the confidence intervals all overlap except for Section B marked with CMI⁺, where the upper limit of 0.490 is outside of, and falls short of, all the other ranges. The t test of the difference between the paper and CMI⁺ correlations is not significant for Section A (t=0.57, df=8, p=0.584) or for the total marks (t=2.16, df=8, p=0.063), but it is significant for Section B (t=7.52, df=8, p=0.000). Thus for Section A there is an equivalent level of agreement between the Principal Examiner and the other examiners whether their marking is on paper or with CMI⁺, but this is not the case for Section B. Table 2 therefore suggests a significant loss in this agreement, i.e. in marking reliability, for electronic compared to paper-based marking of Section B.

Table 2 Coefficients of correlation with Principal Examiner's section and total marks

(a) Paper-based marking (N=35)

	mark allocation	Examiner					mean	95% CI range
		SE1	SE2	AE1	AE2	AE3		
Section A	27	0.831	0.777	0.599	0.787	0.753	0.759	0.673 – 0.824
Section B	27	0.602	0.657	0.716	0.775	0.675	0.690	0.585 – 0.791
Total	54	0.762	0.756	0.716	0.820	0.735	0.760	0.675 – 0.825

¹ The overall correlation for Paper 1 in the earlier study was 0.95, although it is not directly comparable, as will be discussed later.

(b) Electronic marking with CMI+ (N=40)

	mark allocation	Examiner					mean	95% CI range
		SE1	SE2	AE1	AE2	AE3		
Section A	27	0.753	0.721	0.679	0.806	0.692	0.734	0.648 – 0.801
Section B	27	0.292	0.326	0.378	0.433	0.345	0.356	0.205 – 0.490
Total	54	0.609	0.685	0.591	0.798	0.613	0.668	0.516 – 0.749

The disappointing correlations for Section B gave reason to be cautious in moving forward with marking essays on screen, at least with the annotation tool available to examiners at the time of this investigation. This outcome will be discussed further in the light of the examiners' comments.

Examiners' views

Examiners reported adapting to electronic marking with CMI+ and the annotation tool quickly, and finding the procedures straightforward. The conventions used in paper-based marking were quickly adapted to the new medium. For example, ticks (on paper) were replaced with underlining of key words or phrases (on screen), and this was considered a good alternative because it was more specific about what was being rewarded. Marks were placed at the appropriate points in the response. Other annotation options include placing a box around text, and sidelining. Comments and marks on screen appear are in colour and are neat in appearance, but the examiners asked that they be made bolder, to be as prominent as their handwritten comments on paper. Referring back to annotations was held to assist in establishing an objective and replicable mark for the level-based marking of the Section B essay in particular. This agrees with a 'facilitation' as well as a 'justification' function attributed to annotations in marking by Crisp and Johnson (2007), who draw on the work of Bramley and Pollitt (1996) to suggest that *'annotating might reduce the cognitive load of markers during the judging process by creating a 'visual map' of the quality of an answer . . . assisting comparisons with other answers'* (p945).

As already noted, inserting a comment rather than just a mark is time consuming. In addition there was a view that making an annotation interfered with the process of reading the response and placing it at the correct level against the descriptors of performance in the mark scheme. It also became something of a chore to add comments, which are fairly predictable and repetitive because the examiners are looking for evidence for a particular descriptor of performance and noting where they have seen it. This led to a suggestion from one of the senior examiners that electronic marking might use a customised set of shortcut keys to trigger the phrases used from a comment bank. Crisp and Johnson (2007) demonstrated that, within a subject, a common or subject-specific set of annotations is used. Using a comment bank would be less intrusive to the marking process and would also save time. The examiners admitted a tendency to hesitate and then not bother to add a comment (but with assurances that had it been 'live' marking the comment would indeed have been added).

There was some concern that e-marking at home could be more easily disrupted than paper-based marking, especially for the long essay question. This is because for operational reasons the e-Marker® applications 'time out' after a fairly short period of inactivity, and the current response being marked then disappears and is no longer available for marking. Given the time needed to mark each essay, timing out could mean significant costs, whether to the examiner in 'losing' a response after spending time in initial marking, or to the candidate if the quality of the marking suffers through being rushed to completion.

There were rather low expectations for marking the single essay question in Section B electronically, and indeed the analyses suggest that marker reliability may have suffered. The examiners were least happy with marking this question, and some found scrolling up and down through the response on screen less satisfactory than reading it on paper. As a result they were not fully confident that the marks they had awarded were correct, or that on a second viewing they would come up with the same mark. The shorter responses given to the Section A part questions were found easier to deal with on screen.

There was some divergence in views in relation to the 'segmented' or question by question marking of CMI⁺ as opposed to the whole paper marking with which examiners are familiar. The former denies the opportunity to build up a picture of the candidate, and to some this was problematic and not welcome. The alternative view was that segmentation makes the task less demanding. Proponents of both perspectives claimed greater accuracy for their preferred approach.

In summary the examiners' comments suggest that scrolling through a response, locating suitable points to insert comments, and pausing to make annotations via the keyboard, combined with segmented rather than whole paper marking, could all interfere with the processes they would normally use to determine the appropriate level of a response and the appropriate mark to award within that level.

DISCUSSION AND CONCLUSION

The procedure adopted in this investigation diverges of necessity from the reality of 'live' electronic marking. This has implications to consider against the suggestion in the findings of apparent interference with the examiners' normal essay marking processes, detracting from the reliability and possibly the validity of their marking. In an experimental study of this kind a number of the recognised advantages of e-marking are not in evidence but it is appropriate that they be briefly mentioned.

The most obvious differences from the reality of e-marking concern its real-time operational facilities to monitor and enhance the quality of marking; directly comparing the paper and electronic marks ignores the great potential of e-marking to improve marking reliability through live and more sensitive monitoring of individual examiners' marking than is available in conventional paper-based marking. For example, doubtful e-marking can be identified and the examiners given feedback so that they can adjust their marking, or alternatively they can be switched to the questions that they are most competent to mark. This facility will prove especially useful for long written answers to more specialised questions.

As already noted, electronic marking with CMI⁺ raises the issue of differences between whole paper and 'segmented' or question by question marking. Here the GCSE English examiners took the view that they become 'familiar' with a candidate, and not just their handwriting, through their Section A responses. Some claimed that whole paper marking helps in reading, comprehending and assessing the Section B essay and therefore gives more accurate marking. The alternative view is that segmented marking is less demanding because it recognises the 'sharing' aspect of e-marking, where there are typically as many examiners marking an individual script as there are questions on the paper, and avoids any 'halo' effect operating. This aspect could not be incorporated into the design of this investigation, because all the participants marked all the candidates' responses. As a result marking reliability with CMI⁺ and annotation may be under-estimated here.

In segmented electronic marking each response must be associated with a specific question, which is most easily achieved by defining the response areas for candidates to use for their answers. The full effect of this could not be seen in this investigation because existing examination scripts were used, but if pre-defined spaces for responses are introduced for the purpose of segmentation this is likely to influence the amount, and the content, of what is written; candidates will be influenced by how much space is allocated for their response and are likely to treat it as a clue to what is expected (Crisp, 2008), just as they know to treat the number of marks allocated to each question printed on question papers as a pointer to how much 'worth' an answer will be given. Performance might therefore be expected to improve.

Another restriction in this exercise is that it was possible only to involve a single GCSE subject, which restricts the generalisability of the findings to the extent that approaches to essay marking vary from subject to subject. However, like English, they mostly use a 'levels of response' approach to marking. Often they identify a number of sub-headings, each with levels of response descriptors, which it is speculated can make the task of marking long written answers more manageable. The alternative to identifying levels of response is a points-based approach. Where this is used, in a highly specified mark scheme, the facility offered by the annotation tool of automatically adding up the marks awarded at various points in an extended response would doubtless prove very welcome to examiners, and to the awarding body in eliminating errors of addition.

Another limitation was that time constraints meant that while all the electronic marking was carried out in office conditions, this was not true for all the paper-based marking, which the examiners finished off at home, having completed a proportion of it in regular breaks from the screen. Completing at least some of the paper-based marking under office conditions was a deliberate attempt to match the physical conditions of the paper-based and electronic marking, but in so doing the investigation ran the risk of emphasising differences from the marking the examiners normally engage in, and perhaps weakening the reliability of their paper-based marking, despite four of the six examiners being familiar with office-based script marking from their experience of 'mopping up' unmarked scripts in the late summer period before results day.

In comparison with the earlier investigation in GCSE English (Fowles, 2006), where a very high measure of agreement was found for this particular Higher tier paper, there is a key difference in the task presented to the examiners participating in this investigation. The earlier exercise compared live marks originally awarded by assistant examiners in the normal way with the 'true' marks of the Principal Examiner, and, although it was a few months after the examination, the Principal is accustomed to marking many scripts in the post-examination re-marking period. Here *all* the examiners were marking the cleaned scripts, and over a year after the examination. Assistant examiners do not have experience of post-awards re-marking. Although the artificiality of the marking in the current exercise 'cancels out' because it is common to both the paper and the electronic marking, it might explain some of the lower marking reliability for this paper (as indicated by the lower correlation coefficients in Table 2) than would have been expected from the earlier study, coupled with the fact that extra time had elapsed during which the personnel had all been engaged with the following year's examination.

A further limitation relates to the status of the Principal Examiner's marking. Under the hierarchical definition of the 'true' mark, both this and the earlier investigation of marker reliability in GCSE English use the mark awarded by the Principal Examiner as the 'true' mark (Meadows and Baird (in press)). In the present exercise the Principal Examiner provided sets of marks on paper and electronically which were compared with the other individual examiners'

marks. The analysis reported above therefore rests on the assumption that the same relationship can be expected between each examiner's marking and that of the Principal Examiner in both marking contexts (paper and electronic). The Principal Examiner was of course no more familiar with electronic marking than the other examiners, and this makes the assumption a little tenuous: any reservations about the electronic marking of this paper, and of Section B in particular, apply as much to the Principal Examiner's marks as to the examiners' marks, with the implication that an additional source of marking unreliability has been introduced for electronic marking in the design of the study because of increased uncertainty as to the 'true' mark. One solution for investigating marking reliability in a novel context such as e-marking would be to relieve the Principal Examiner's marks of the status of 'true' marks in favour of a consensus definition of 'true' marks, although such alternative 'true' marks would have to be generated from a much greater number of markers than in the present study (and that would never be feasible operationally).

In conclusion, the analyses comparing examiners' marks relative to those of the GCSE English Principal Examiner in the two contexts of paper-based and electronic marking found few statistically significant differences other than in the coefficients of correlation, which were lower electronically than on paper for the essay section. This might suggest that the annotation feature itself has caused a disturbance in marking reliability of the essay in Section B that is not in evidence for the shorter prose responses of Section A. The evidence is not conclusive, not least because of the various necessary limitations of the investigation discussed above, but the participants were certainly less confident in their marking of Section B, and it is suggested that scrolling through a response, locating suitable points to insert comments, and pausing to make annotations via the keyboard, combined with segmented rather than whole paper marking, all interfered with the processes by which they normally determine the appropriate level of a response and the appropriate mark to award within that level. It is a major concern for AQA that developments in technology for assessment purposes, that are expected to introduce gains in efficiency and marking reliability, do not instead detract from reliability or call the validity of the assessment into question. Each development is therefore introduced cautiously and incrementally, at first to a limited range of examination components. It can however be noted here that the exercise suggested some beneficial modifications to the annotation tool that have been taken further, with some already implemented. It is also worth noting that e-marking has facilities for real-time monitoring that could not be exploited in the design of this re-marking investigation, but can firmly be expected to enhance marking reliability.

Dee Fowles
AQA Research and Policy Department
e-mail: dfowles@aqa.org.uk

REFERENCES

- Bramley, T. & Pollitt, A. (1996) *Key Stage 3 English: Annotations Study*. A report by the University of Cambridge Local Examinations Syndicate for the Qualifications and Curriculum Authority (London, QCA), cited in Crisp and Johnson (2007).
- Crisp, V. (2008) Improving students' capacity to show their knowledge, understanding and skills in exams by using combined question and answer papers. *Research Papers in Education*, v23 n1 p69-84.
- Crisp, V. and Johnson, M. (2007) The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*, v33 n6 p943-961.
- Fowles, D. (in submission) How reliable is marking in GCSE English?

- Fowles, D. and Adams, C. (2005) *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the IAEA Conference, Abuja, Nigeria.
- Hewitt, E.A. (1967) *The reliability of G.C.E. O-level examinations in English Language*. Occasional Paper 27, Joint Matriculation Board. Manchester: JMB.
- Shaw, S. (2008) Marking essays on screen: towards an understanding of examiner assessment behaviour. *Research Matters: A Cambridge Assessment Publication*, 6, p9-15.
- Johnson, M. and Shaw, S. (2008) Annotated to comprehend: a marginalised activity? *Research Matters: A Cambridge Assessment Publication*, 6, p19-24.
- Meadows, M. and Baird, J. (in press) What is the right mark? Respecting other examiners' views in a community of practice.
- Meadows, M. and Billington, L. (2005) *A Review of the Literature on Marking Reliability*. Unpublished AQA report produced for the National Assessment Agency.
- Raikes, N, Greatorex, J. and Shaw, S. (2004) *From Paper to Screen: some issues on the way*. Paper presented at the IAEA Conference, Philadelphia, USA. Retrieved 11 July 2008 from <http://www.cambridgeassessment.org.uk>.
- Spearman, C. E. (1904a) 'General intelligence' objectively determined and measured. *American Journal of Psychology*, v5, p201-293.
- Spearman, C. E. (1904b) Proof and measurement of association between two things. *American Journal of Psychology*, v15, p72-101.
- Spearman, C. E. (1927) *The abilities of man, their nature and measurement*. New York: Macmillan.
- Times Online (2008) *Markers award students for writing obscenities on GCSE papers*, 30 June 2008. Retrieved 11 July 2008 from http://www.timesonline.co.uk/tol/life_and_style/education/article4237491.ece.