LITERATURE REVIEW ON EFFECTS ON ASSESSMENT OF E-MARKING

Diana Fowles

Abstract

The development of e-marking is reviewed under the headings of the different technology suppliers that are driving e-marking developments forward through their software solutions. The review covers the experience of AQA and other UK awarding bodies, and considers the various aspects of e-marking, a hybrid in which examiners carry out on-line marking of extracts from scanned scripts, which make it different from conventional marking. These include marking single items rather than whole scripts and allocating some items for marking by experts or generalists and others for 'automatic' marking. Such differences might conceivably influence the examiner's response to the task, and possibly the marks awarded also. Although any impact on the assessment is seen as much less significant than might be found in developments involving on-line testing and full computer marking, a few issues are identified for the validity and reliability of the assessments delivered with e-marking which might be investigated further.

1 Background

1.1 Consideration of the potential of e-marking of candidates' responses in scanned or imaged form on computer screens, rather than in their original paper format, has tended to focus on structural and administrative aspects, understandably given concerns about the costs of examining and escalating demands on examiner recruitment. There are numerous practical questions in moving to more technologically driven systems. AQA's first two live operations[1] using Computer Marking from Image (CMI+) provided by DRS Data and Research Services plc have been initiated early in 2005 ahead of the Summer 2005 series when it is planned to e-mark 28 GCSE examination components or modules. What is less frequently considered in moving forward with these developments is the impact on examining of performing the e-marking task in place of traditional paper-based methods.

1.2 At its meeting on 1 July 2004, the AQA Council considered a paper on "The Business Case for E-marking and E-mark capture". The main focus of the paper was to summarise the different initiatives in this area, discuss their operational benefits and present an overview of the costs involved, with more detailed cost/benefit analyses being undertaken by the Finance Department in conjunction with the Business Development Department and DRS.

[1] Both involve module tests: the January 2005 French Specification B Module 1 Listening test and the March Mathematics Specification B Module 3 Foundation Tier test.

1.3 Although the paper focused on operational and financial aspects, there was also some discussion about the impact of introducing these technologies, especially on-line marking (e-marking), on the characteristics of the assessments made. A discussion arose not just about how CMI+ could most easily be implemented by AQA, but what the assessment implications were of so doing. Initial discussions with the New Technology Co-ordinator indicated that, for the next few years at least, there would be minimal design (and hence cost) implications for conventional question paper design. There would, however, be much to gain in terms of reduced examiner fees - plus improvements in reliability and efficiency - a view shared by AQA's technology partner DRS.

1.4 The Council suggested that a literature review be conducted on the assessment/educational implications, of introducing e-marking, especially with respect to validity, lest they seriously undermine and jeopardise the quality of the assessments made by the AQA, and the public confidence they currently demand.

1.5 This review is concerned with e-marking, where human markers judge candidates' work displayed on a computer, particularly as it is developing through CMI+. It is not concerned with e-capture of marks or the broader area of e-assessment (also known as computer based assessment (CBA) and computer assisted assessment (CAA), and comprehensively reviewed by Ridgway and McCusker (2004)). The former (e-capture), with which AQA has now gained some experience (Arlett and Bridge, 2004; Arlett, 2004; Fowles, 2004), should have no implications for current assessment methods. The latter (e-assessment) presents the assessment task to the candidate by computer and is a future rather than immediate prospect for AQA, with mostly separate assessment implications.

2 Introduction to the review

2.1 A pre-requisite for an assessment to be valid is that it must have reliability, while a reliable assessment has meaning only if it also has validity. If tests are re-designed or modified in any way to make them more suitable for e-marking there is a danger that validity could be an issue, for example if the lower costs of automatic and clerical marking were to introduce a shift towards a greater proportion of items that can be marked in this way.

2.2 Below are a number of the questions related to validity and reliability that might be relevant to the move from conventional to e-marking of an examination component, and this review looks for evidence that might help in considering them.

VALIDITY

(a) Is there any evidence of assessment schemes being developed and shaped by what technology can offer, at the possible expense of the validity of the assessment?
(b) Is the validity of the assessment retained when it is carried out, at question or part question rather than whole paper level?
(c) Is the validity of the assessment retained when it is e-marked by experts and generalists, or by the computer, depending on the complexity of the marking?

RELIABILITY

(d) Do candidates' total marks differ when e-marking is of a *whole component* on-screen (as in CMI) rather than conventionally?

(e) Are candidates' marks affected by marking being carried out *at question or part question level* (as in CMI+) rather than at whole paper level (including such factors as examiner accuracy, halo effects, over-penalising of repeated errors). If so, which is the more reliable? Are any differences in reliability the same for components made up of short answer as opposed to longer, free response responses?

(f) How important to the reliability of marking is the facility to annotate scripts when marking conventionally? How satisfactory are computer-based annotations and comments?

(g) A related but not identical question is whether reliability is enhanced if markers can revisit questions/papers they have already marked, whether to amend the mark given or simply to view?

(h) What are the implications for the reliability of assessments overall if examiner satisfaction is reduced when expert examiners mark electronically rather than conventionally?

(i) Are there any implications for reliability of marking if the pool of examiners for a particular component/subject loses examiners who are not technologically well enough equipped or willing to e-mark? Are they likely to be randomly distributed or unrepresentative in terms of their reliability of marking?

(j) Will there be problems in maintaining a pool of expert examiners for each component if the number of expertly marked responses varies significantly from series to series, and if so will this affect reliability?

2.3 By the nature of the technology required to support e-marking, literature that might address all bar one of these questions will have been produced within the past five years or so. The exception is the part versus whole paper question ((b) and (e)), although it has not been an issue of great concern in relation to conventional marking. Either approach has been used, often according to whichever has been found to be more efficient or administratively easier. In Higher Education for example the tendency is to mark questions individually (unavoidable when different markers take responsibility for different questions), while school examination papers have usually been marked as whole papers, although individual markers can elect to mark question by question within their script allocations if they so prefer. The two approaches have continued into on-screen marking, where marking can be of whole papers (e.g. in the CMI development) or of papers split into parts, usually single questions (as in CMI+). These approaches are returned to later (para. 4.1).

2.4 Most attention in the literature is given to CBA/CAA, i.e. to initiatives designed to move away from paper to a wholly digital assessment system, in which candidates respond directly to the computer. Within CBA the marking can be on-line, by human markers, but the vision is of systems where the computer marks the answers, and delivers the test result. Advances in technology mean that 'the infrastructure is quickly falling into place for Internet delivery of assessment to schools' (Bennett, 2001). Interest and effort appear greatest for computer testing combined with computer marking, and not only in the US – in this country, for example, the annual CAA Conference, hosted by Loughborough University, attracts a wide range of potential and actual users, especially

from Higher Education. AQA's experience in this area so far is limited to Key Skills testing. The on-line marking by examiners of scanned scripts can be seen as a hybrid system which is "a way of realising some of the benefits of digital scripts in a context where paper is likely to remain important for many years to come" (Raikes, Greatorex and Shaw (2004)). In consequence the literature is rather sparse for the development of e-marking, which has found particular favour in the UK examination context, with little from American sources.

2.5 In an OCR paper presented to the IAEA conference in June 2004, Raikes *et a*l provide both a review of the existing literature on e-marking in place of conventional marking, and a long list of questions and issues for research raised by the move to e-technology using scanned scripts and e-marking. The three authors report on OCR's experience of an e-marking trial in January 2004, particularly on the reactions of the examiners taking part. Their review of the existing literature includes only a few non-UK authors, all from the Educational and Testing Service (ETS), and all writing on experience with the Online Scoring Network (ONS) software developed by ETS whose software OCR was then using.

2.6 There are other terms for what AQA and others refer to as 'e-marking', including on-line marking, on-screen marking, on-line scoring, etc. Here the term 'e-marking' will be used exclusively, even where the process is known by one of the alternative terms in the context being described.

3 Review papers

3.1 This review is organised under the headings of the particular technology supplier driving the developments forward, starting with experience with ONS and moving on to experience gained in the UK with the two other major e-marking software initiatives.

3.2 ETS's Online Scoring Network: ONS

The first ETS paper is by Zhang *et al* (2003), on using ETS's ONS software with the Advanced Placement Program testing (AP). ONS was developed to accommodate marking away from centralised locations via the Internet to remote locations, i.e. for home marking. Some of the research reviewed by Zhang *et al* is on the impact of this change, which is not of course an issue where UK examinations are concerned. Other ETS research reported in this paper was carried out by Powers and associates (1997 and 1998) can be summarized very briefly: when experienced readers marked essay responses on paper and using ONS, there were no differences in average scores awarded in either medium. The other ETS source identified by Zhang *et al* is Bennett (2003), who also found that "the available research suggests little, if any, effect for computer versus paper display". As for the AP tests, some 6,000 examination papers for two tests were marked with ONS and 'the results obtained with ONS were found to be extremely similar to those obtained with traditional AP scoring methods'. Altogether these studies gave ETS reassurance as to the feasibility of e-marking, and ONS is seen within ETS as an important IT development 'for organising the (marking) process itself and enhancing its efficiency and quality' (Bakker and van Lent, 2003).

ONS software was also used by AQA in a marking centre trial involving four examiners from KS3 English and three from KS2 Mathematics (Royal-Dawson, 2003). The trial

included some double marking (12%) as a reliability check, and markers were also required to mark 100 scripts conventionally for a comparison of ONS and conventional marking. For English the conventional total mark was higher than the e-mark total on average by 0.60 out of 17 (3.5%) whereas for Mathematics the mean difference was negligible (0.02 out of 20). The report on the trial makes clear that the question of the reliability of e-marking remains, but that in conjunction with the findings of other studies, there is a suggestion that there is very little difference between e-marking and conventional marking for tests made up of objective type items, whereas for tests with items that require a considered judgement, conventional marking yields relatively higher marks. One explanation may lie in the part versus whole approach to the marking in the two modes, and will be considered further later.

The OCR e-marking trial described by Raikes *et al (op cit)*, which involved four GCE subjects (mathematics, physics, chemistry and general studies), also used ETS's ONS software. Three senior examiners plus the appropriate Chair of Examiners for each of the four subjects were brought together for a day's trial in the UCLES offices. They noted various practical issues that were of concern and made suggestions for improvements, for example, to be able to make textual comments when referring scripts to a Team Leader.

The present reviewer has not found any more recent publication concerned with e-marking than that of Raikes *et al.* Many of the questions they pose relate to the quality of assessments made using e-marking and have still to be fully answered, including the following.

- What question paper and answer booklet designs are most effective at encouraging candidates to write correctly labelled answers in appropriate places, using appropriate materials?
- What changes to question papers and mark schemes might facilitate marking by non-expert examiners, including automatic methods i.e. marking by the computer?
- Can all existing paper-based examinations be marked on-screen or are there features of question papers which cannot be accommodated or which are too costly to accommodate?
- What training and standardisation methods are most appropriate for different marker groups, i.e. general markers and expert examiners?
- Does standardisation using paper scripts transfer to marking on screen?
- What feedback should be given to markers? Can feedback during marking lead to undesirable disturbances in markers' behaviour?
- Should Principal Examiners mark whole scripts, even if other markers do not, as the basis for recommending grade boundary marks to the awards meeting? If they do not mark whole scripts, can they make judgments about boundary marks by looking at scanned scripts with a print-out of the marks awarded to each question (the method used in AQA awards meetings?)

OCR next trialled DRS's e-capture software (OMS, QMS and CMS) using five CIE components early in 2004 (UCLES, 2004), subsequently moving on to work with Research Machines plc as its lead technical partner from Summer 2004, and returning to developments in e-marking.

3.3 Pearson's Electronic Performance Evaluation Network: 'e-pen'

In the UK the published evidence on e-marking is mostly from NFER and National Curriculum testing, using NCS Pearson's 'e-pen' (or alternative Netgrade software) and centralised marking. Trial e-marking of Year 7 Progress tests by both 'clerical' and 'expert' markers was evaluated by Newton *et al* (2001) and Whetton and Newton (2002). The non-expert marking for questions designated as not needing experts was found to be effective. Marking at home using the Internet instead of central marking was seen as the next necessary development. Some differences in results from conventional marking and e-marking were found in two categories of the testing (writing and spelling/handwriting). Sturman and Kispal (2003) also compared conventional with e-marking of optional English tests designed for pupils in Key Stage 2 (reading, writing and spelling, at the pre-test stage) and found no consistent trends in the differences. They suggested that their results were likely to be more typical than those of Whetton and Newton, and called for further studies to look for differences.

AQA also gained e-marking experience using NCS Pearson's software with GCE Chemistry papers. Unit 1 examiners were invited to participate in non-live marking on the company's premises. Some questions were identified for clerical[2] marking, and their marks and those of the expert examiners were compared with the live marks. There were small differences for individual items but in both directions, so that overall the mean difference in total marks was only 0.13 marks out of the total of 90 (Fowles, 2002). A clear preference was expressed by the expert examiners for being able to e-mark at home. The examiners noted various practical issues that to a great extent have been resolved in the DRS software of AQA's recent trials.

Edexcel uses Pearson's software[3] but, although it provides speakers to contribute on the topic of e-marking to conferences e.g. the Association of Colleges Conference in November 2004, it has not published in this area, presumably because of commercial sensitivity and lack of research capacity. Working within Pearsons, Twing, Nichols and Harrison (2003) compared results for a large-scale writing assessment in the US, in which essays were marked (each out of 6 marks) both conventionally and with ONS scanned images. They report slightly more reliable and more accurate marks for ONS although the differences were 'small and not practically meaningful'. A marking rate analysis suggested that e-marking could be more than 15 per cent faster.

3.4 DRS's Computer Marking from Image: CMI and CMI+

The business case report provided to the AQA Council in July 2004 gave a history of the e-marking work carried out by AQA to that date, outlining the benefits of the e-marking technology (Appendix) and recording the decision to use DRS's CMI+ system. QCA staff were present to observe AQA's next step, of live CMI+ e-marking of the January 2005 GCSE French Specification B Module 1 Listening test, with centralised general and expert e-marking in Harrogate. QCA had raised a number of questions arising from the changes that accompany e-marking, of the kind posed by Raikes *et al*,

[2] 'Clerical' marking is now referred to within AQA as 'general' marking.

[3] There is little information on the Edexcel website, while the new Pearson website appears to have dropped some evaluative articles on e-pen which were previously available on the NCS Pearson website.

recorded earlier. No particular worries about the transfer from conventional to e-marking arose in Harrogate. Automatic, general and expert marking were all included, with more than half of the responses being automatically marked. On this first live e-marking occasion, as a one-off, all responses for human marking were double-marked and the response to QCA was able to report an extremely high level of agreement between markers, both general and expert, with differences for only 1.6% of responses (Fowles, 2005). Responses to items identified for automatic marking are all routinely double-keyed. The responses are then all listed with their frequencies and presented to the senior examiner, whose task is to mark each response on the list. The computer then 'marks' each candidate's response, i.e. it allocates the mark determined for that response in the senior examiner's marking rules. Automatic marking is thus perfectly reliable in the sense that it will produce the same set of marks on a second occasion of marking, although a second set might differ if a second examiner were to provide the marking rules.

WJEC has also embarked on a programme of work with DRS, using both e-capture through CMS and also CMI. CMI does not depend on Internet, particularly broadband, access (a difficulty for rural parts of Wales), instead it allows for complete scanned scripts to be downloaded to CDs, for distribution to examiners to mark at home. One of the reasons AQA opted for CMI+ was that lower costs were anticipated through being able to direct some items for marking by general markers and others for automatic marking by computer. Using CMI does not offer these savings.

The evaluation of the WJEC pilot in 2004 identified a difficulty for senior examiners in using CMI (but not shared with CMI+), which is that they have to mark assistant examiners' sample scripts blind. This results in a departure from their conventional marking where they see the marks awarded by an assistant as they mark, and make a professional judgment about the appropriateness of that marking. Not surprisingly it was found to introduce more variation between supervising and assistant examiners than had historically been the case with the e-marked component (a GCSE ICT paper) (WJEC, 2004).

4 Research Issues

4.1 Part versus whole paper marking

Although segmentation, or part versus whole marking, is a topic that might be expected to have received attention in the wider assessment literature, little reference to this aspect of marking has been found. In conventional marking, examiners mark complete scripts within their allocations and, although specific instructions are not given on how the marking is approached, it is reasonable to assume that in most cases it will be of each paper as a whole – if only because the format of presentation of the responses in answer booklets makes this the most convenient and efficient way to carry out the task.

Scharachskin and Baird (2000) reported on how consistency of performance (or lack of) across an examination paper can affect grading decisions when a whole paper is reviewed after marking has been completed according to the mark scheme. They concluded from observations in Biology and Sociology that appreciation of consistent performance throughout a paper was 'a feature of the examination performance that was not part of the marking scheme (but) affected grading decisions' when examiners

were asked to grade scripts with borderline marks. They suggested that there could be 'tunnel vision' effects when examinations are broken into small parts for marking (Baird and Scharachskin, 2002).

In an earlier OCR paper on e-marking pilot work, Raikes (2002) reported that one of the examiners' main concerns was 'a perceived need to know a script's total mark …. some English Literature examiners said that they needed to mark a whole script to award a fair mark'. Sturman and Kispal *(op cit)* commented that in whole-paper marking 'it is theoretically possible for (markers) to build up a 'picture' of each pupil's attainment as they mark'. They did not however find support for this suggestion in their analyses of questions set on a narrative (story) in National Curriculum reading tests for pupils in Key Stage 2. They speculated that a picture of the pupil would be most likely to build up in these questions but, in each of the three age groups tested, the e-marking narrative scores were actually higher than the conventional scores.

The view that segmentation in marking is potentially more objective than marking of the whole paper (or at least no less objective), has informed AQA's readiness to accept question level marking. As e-marking is extended there will be more opportunities for empirical study of the view that segmentation can ''add to the objectivity of the marking' (Bakker and van Lent, *op cit*). Williams and van Lent (2002) identified three particular factors expected to contribute to the fairness of e-marking of parts: (a) the complete anonymity of the responses being marked (the items being marked carry no name, gender or school information); (b) minimal opportunity to build up a 'halo' effect', where early answers influence scores later in a test; and (c) the random allocation of a candidate's responses to a range of markers, which means that any examiner error in marking, whether from marking severely, generously or erratically, will be randomly distributed across individual candidates. This last factor means that mark/re-mark reliability should be higher than if one marker had marked all the items.

The isolated effect of segmentation on reliability might be investigated if a sample of scripts were to be marked by a sample of examiners (on paper, not electronically), first with the responses copied on to separate sheets of paper, and presented for marking question by question, and second as whole papers.. Such an exercise would provide two sets of item and total marks from paper-based marking, and any consistent differences looked for. The exercise might be repeated for electronic marking (i.e. using CMI and CMI+). Previous work cited earlier has compared *conventional* total marks with *e-mark* totals, but what is suggested here would compare totals from within the same medium (paper or electronic).

Raikes (*op cit*) noted that Mathematics examiners found item level marking 'boring or less rewarding than marking whole scripts' and this observation suggests that there may be differences between subjects in readiness to move from whole to part marking.

This last observation suggests that there will be different responses to e-marking from different groups of examiners, which AQA research will need to monitor, as it has with the introduction of the different forms of e-capture (Arlett, Arlett and Bridge, Fowles, all *op cit*).

4.2 Script annotation

In some subjects examiners are used to annotating scripts. Some examples are given by Raikes *et al*, e.g. 'ECF' in Mathematics for 'error carried forward' (to indicate that a further mark has not been deducted). They note that examiners have reported feeling that they need to be able to annotate scripts as they e-mark 'in order to mark properly'. This may only mean putting in ticks, but it appears to be a feature of conventional marking that serves an essential function, reassuring to the examiner while marking and valuable should the marking be challenged later. Similar comments have been reported from moderators in relation to e-portfolios (Greatorex, 2004).

Annotations may influence the process of marking and thereby its reliability. One specific function of script annotations is to meet aspects of the mark scheme that are not tied to any particular question, for example the marks for spelling, punctuation and grammar, or quality of written communication.

DRS was able to add a 'comment' facility for the January 2005 French live e-marking, which was well used, and at last partly removes a difference between the two forms of marking. It is most likely that technology will further develop to allow examiners to make whatever annotations and comments that wish, according to their current practices.

4.3 Comparing responses and amending marks

Another aspect of conventional marking that examiners have requested for e-marking is a facility to return to previously marked responses, in order to compare them and adjust marks accordingly. This feature is also important in the e-portfolio context where moderators need the facility to call up work they have already seen for further review (Greatorex, *op cit*). This is again a feature of marking that may influence its reliability. Mark corrections need sometimes to be made – examiners report inadvertently entering an incorrect mark when e-marking and needing to correct it (Fowles, 2005). This presents a problem given that the responses are not subsequently identifiable by the examiner, but hopefully software solutions will be found for CMI+.

4.4 Validity of expert / general / automatic marking

Williams and Bakker (2001) at ETS analysed Key Stage 2 question papers in English, Mathematics and Science and reported to QCA that most were suitable for e-marking, and in the form in which they were written. They reported that a proportion could be marked automatically, with varying degrees of change which would mostly be only to layout and presentation, probably therefore without interfering with their 'assessment goals'. Benson (2002) drew on AQA's experience with both ONS and e-pen to consider in some detail the suitability of different types of GCSE questions for e-marking, whether by experts, generalists or automatically. He notes that adapting questions to make them more suitable for e-marking, e.g. to ensure a manageable scan area without too much scrolling around the screen, can potentially change the validity of the questions. In a different format the response to a question may differ and the question could 'work' differently in terms of its facility and discrimination characteristics. An example of a minimal change that would probably not detract from validity is a question type that asks candidates to put a ring around their choice of the correct alternative from

a number offered. A candidate might respond with too large a ring for the defined scan area for the question, and this danger might suggest changing the format, e.g. by giving boxes to tick.

Validity issues might however be raised if e-marked question papers were commissioned to have a pre-determined balance between expert, general and automatically marked questions, perhaps to reduce costs, or simply to ensure a consistent workload for expert examiners. If this were to be the case it would add an extra dimension to question paper preparation which does not need to be considered for conventional paper-based marking.

5. Conclusion

Various aspects of e-marking make it different from conventional marking, and could introduce concerns for the validity and reliability of the assessments made. Although such concerns should be investigated further as more components are drawn in to e-marking, such empirical evidence as is available does not suggest that there are any concerns of such significance that they would cast doubt on the course of action AQA is taking in introducing e-marking. Introducing e-marking by examiners of scanned scripts is a hybrid system that is much less of a sea change than computer based developments involving on-screen testing and computer marking, which are given much greater prominence in the literature.

Dee Fowles
AQA Research and Policy Analysis
March 2005
Revised April 2005

References

Arlett, S. J. (2004) *Electronic Capture of Marks using DRS Software Trial based on AQA GCSE Mathematics Module 1, March 2004 Series* AQA Research Committee paper RC/282

Arlett, S. J. and Bridge, N. (2004) *Electronic Capture of Marks using DRS Software* AQA Research Committee paper RC/275

Baird, J. and Scharaschkin, A. (2002) Is the Whole Worth More than the Sum of the Parts? Studies of Examiners' Grading of Individual Papers and Candidates' Whole A-Level Examination Performances. *Educational Studies, Vol. 28, No. 2, 143-162.*

Bakker, S. and van Lent, G. (2003) *National Testing On Line How Far Can We Go?* Paper presented at the IAEA Conference, Manchester, retrieved 11 February 2005 from http://www.aqa.org.uk/support/iaea/papers.html

Bennett, R.E. (2001*)* How the Internet Will Help Large-Scale Assessment Reinvent Itself*. Education Policy Analysis Archives, Vol. 9, No. 5*

Bennett, R.E. (2003*) On-line assessment and the comparability of score meaning*. ETS Report RM-03-05, presented at the IAEA Conference, Manchester, retrieved 11 February 2005 from http://www.aqa.org.uk/support/iaea/papers.html

Benson, M. (2002) *Impact of E-marking.* AQA report to QCA, Project 2F.1: Impact of e-marking on test design

Fowles, D.E. (2002) *Evaluation of an e-marking pilot in GCE Chemistry: effects on marking and examiners' views.* AQA Research Committee paper RC/190

Fowles, D.E. (2004) *Electronic Capture of Marks using DRS Software: OMS, QMS and CMS Trials, Summer 2004.* AQA Research Committee paper RC/282

Fowles, D.E. (2005) *Evaluation for QCA of the CMI+ System in the January 2005 e-marking live pilot.* AQA Internal paper.

Greatorex, J. (2004) *Moderated e-portfolio project evaluation* . Retrieved 11 February 2005 from http://www.ocr.org.uk/OCR/WebSite/Data/Publication/E-Assessment%20Materials/Moderated_82372.pdf

Powers, D. and Farnum, M (1997) *Effects of Mode of Presentation on Essay Scores.* ETS Report RM-97-08

Powers, D. Kubota, M., Bentley, J.,Farnum, M., Swartz, R. and Willard, A. (1997) *A pilot test of on-line essay scoring.* ETS Report RM-97-07

Powers, D., Farnum, M., Grant, M and Kubota, M (1998) *Qualifying Essay Readers for an On-Line Scoring Network.* ETS Report RM-98-20

Raikes, N. (2002) *On-screen marking of scanned paper scripts.* Cambridge: UCLES

Raikes, N, Greatorex, J. and Shaw, S. (2004) *From Paper to Screen: some issues on the way.* Paper presented at the IAEA Conference, Manchester (retrieved 11 February 2005 from http://www.ucles.org.uk/assessmentdirectorate/articles/confproceedingsetc/IAEA2000NRJGSS

Ridgway, J. and McCusker, S. (2004) *Literature Review of E-assessment.* Report 10, NESTA Futurelab Series, retrieved 24 March 2005 from www.nestafuturelab.org/research/lit_reviews.htm

Royal-Dawson, L. (2003) *Electronic Marking with ETS Software.* AQA Research Committee paper RC/219

Scharaschkin, A. and Baird, J. (2000) The Effects of Consistency of Performance on A Level Examiners' Judgements of Standards. *British Educational Research, Vol. 26, No. 3, 343-5.*

Sturman, L. and Kispal, A. (2003) *To e or not to e? A comparison of electronic marking and paper-based marking.* Paper presented at the IAEA Conference, Manchester, retrieved 11 February 2005 from http://www.aqa.org.uk/support/iaea/papers.html

Twing, J.S., Nichols, P. and Harrison, I. (2003) *The Comparability of Paper-Based and Image-Based Marking of a High-Stakes, Large-Scale Writing Assessment.* Paper presented at the IAEA Conference, Manchester, retrieved 11 February 2005 from http://www.aqa.org.uk/support/iaea/papers.html

UCLES (2004) *Electronic Return of Marks Project.* Internal paper. Cambridge: UCLES

Whetton, C. and Newton, P. (2002) *An evaluation of on-line marking.* Paper presented at the IAEA Conference, Hong Kong

Williams, H. Gray and Bakker, S. (2001) [Project report to QCA]. Utrecht: ETS Europe

Williams, H. Gray and van Lent, G. (2002) [Presentation by Educational Testing Service to QCA, Project 2F.1: Impact of e-marking on test design] Utrecht: ETS Europe

WJEC (2004) *CMI/CMS 2004 Pilot Evaluation*. Internal paper. Cardiff: WJEC

Zhang, Y., Powers, D.E., Wright, W. and Morgan, R. (2003) *Applying the OnLine Scoring Network (OSN) to Advanced Program Placement Program (AP) Tests.* ETS Research Report RR-03-12