Crisp, V. (2007). *Comparing the decision-making processes involved in marking between examiners and between different types of examination questions*. Paper presented at the British Educational Research Association Annual Conference, London.

Crisp, V. (*in press*). Exploring the nature of examiner thinking during the process of examination marking, *Cambridge Journal of Education*.

Crisp, V. (*in submission*). Towards a model of the judgement processes involved in examination marking.

Elander, J. & Hardman, D. (2002). An application of judgment analysis to examination marking in psychology. *British Journal of Psychology*, **93**, 303–328.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: verbal reports as data*. London: MIT Press.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, **19**, 246–276.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision making behaviour of composition-markers. In: M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment*. Cambridge: Cambridge University Press.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. & Gower, R. (1995). *The dynamics of GCSE awarding*. Report of a project conducted for the School Curriculum and Assessment Authority, School of Education, University of Nottingham.

Sanderson, P. J. (2001). *Language and differentiation in examining at A Level*. PhD Thesis. Unpublished doctoral dissertation, University of Leeds, Leeds.

Scharaschkin, A. & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal*, **26**, 3, 343–357.

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? In: L.Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J: Ablex Publishing Corporation.

Marking essays on screen: towards an understanding of examiner assessment behaviour

Stuart Shaw CIE Research

Introduction

Computer assisted assessment offers many benefits over traditional paper methods. In translating from one medium to another, however, it is crucial to ascertain the extent to which the new medium may alter the nature of the assessment and marking reliability. Appropriate validation studies must be conducted before a new approach can be implemented in high stakes contexts. The pilot described here is the first attempt by Cambridge International Examinations (CIE) to mark, on-screen, extended stretches of written text for the Cambridge Checkpoint English Examination. The pilot attempts to investigate marker reliability, construct validity and whether factors such as annotation and navigation differentially influence marker performance across the on-paper and on-screen marking modes.

Candidates wrote their answers on paper scripts in the normal way. The scripts were then scanned and digital images of them were sent by secure electronic link to examiners for on-screen marking at home using Scoris® software.

It can be relatively hard for examiners to make a full range of annotations when marking on screen. For this reason annotation sophistication was manipulated in the pilot as well as marking mode. Four marking methods were compared: on-paper with sophisticated annotations (current practice), on-paper with simplified annotations, on-screen with sophisticated annotations, and on-screen with simplified annotations.

The research literature

There is a large research literature relevant to this project. Key aspects of this literature are summarised below.

Comparability of marking across on-screen and on-paper modes

The literature is mixed on this topic.

- Bennett (2003) carried out an extensive review of the literature and concluded that 'the available research suggests little, if any, effect for computer versus paper display' (p.15).
- Differences were found in a few studies not reviewed by Bennett, however, e.g. Whetton and Newton (2002) and Royal-Dawson (2003).
- Sturman and Kispal (2003) observed quantitative differences between online and conventional marking of tests of reading, writing and spelling for pupils typically aged 7 to 10 years, but an analysis of mean scores showed no consistent trend in scripts receiving lower or higher scores in the e-marking or paper marking: 'absence of a trend suggests simply that different issues of marker judgement arise in particular aspects of e-marking and conventional marking, but that this will not advantage or disadvantage pupils in a consistent way' (p.17). Sturman and Kispal concluded that e-marking is at least as accurate as conventional marking. Wherever differences between the

two marking modes existed they tended to occur when marker judgement demands were high. They also noted that when assessing a pupil's response on paper, holistic appreciation of the entire performance may contribute to a marker's award, but this is not possible if scripts are split up by question for on-screen marking.

- Shaw, Levey and Fenn (2001) have investigated the effects of marking extended writing responses across modes. Scripts from Cambridge ESOL's December 2000 Certificate in Advanced English examination, were scanned and double-marked on-screen. Statistical analysis of the marking indicated that examiners awarded marginally higher marks on-screen and over a slightly narrower range of scores than on paper. The difference in marking medium, however, did not appear to have a significant impact on marks.
- Twing, Nichols, and Harrison (2003) also looked at extended prose on screen. The allocation of markers to groups was controlled to be equivalent across the experimental conditions of paper and electronic marking. Findings revealed that marks from the paper-based system were slightly more reliable than from the screen-based marking. The researchers canvassed opinion from markers and deduced that for some, interaction with computers was a new experience. For these markers, lack of computer experience and familiarity engendered anxiety about on-screen marking. Research suggests that anxiety over computer use could be an important factor militating against statistical equivalence (McDonald, 2002). Mere quantity of exposure to computers is not sufficient to decrease anxiety (McDonald, citing Smith, Caputi, Crittenden, Jayasuriya and Rawstorne 1999) – it is important that users have a high quality of exposure also. Interestingly, for those markers experienced with computers, Twing *et al*. (2003) found that image-based markers finished faster than paper-based markers.
- The question of whether examiners make *qualitatively* different judgements when marking the same piece of writing in different marking modes is a key consideration in assessment (Shaw and Weir, 2007). There is very little research to draw upon in this area. Johnson and Greatorex (2006) conclude that judgements made on-screen and conventionally on paper are qualitatively different, stressing that effects of mode on assessment evaluations are both important and in need of on-going inquiry.
- Although much evidence suggests that examiners' on-screen marking of short answer scripts is reliable and comparable to their marking of the paper originals, it is clear that more research is needed, particularly concerning assessment of extended responses on-screen, to ascertain in exactly what circumstances on-screen marking is both valid and reliable.

Examiners' annotations

- There is a relative paucity of literature relating to the use, purpose and application of annotations in examination marking.
- Crisp and Johnson (2005) suggest that annotations serve two distinct functions: as an accountability function (*justificatory*) and as a means of supporting examiners' decision-marking processes (*facilitation*).

Justificatory function

- Murphy (1979) notes that senior examiners are influenced by the marks and comments on scripts during the process of review marking.
- In their experimental study on the use of annotations in Key Stage 3 English marking, Bramley and Pollitt (1996) observed that 'having annotations on the scripts might enable team leaders to identify markers whose marks need checking' (p.18).
- As part of an investigation into marking reliability involving double marking, Newton (1996) explored whether correlations between first and second marks were affected by obscuring the first marker's comments from the second marker. Newton presented second markers with 'partially obscured' scripts, where the first marker's marks had been obscured but the comments left visible, and 'fully obscured' scripts, where both marks and comments had been obscured. The correlation between first and second marks was a little higher for the partially obscured scripts, but the difference did not reach statistical significance.
- Williamson (2003) asserts that annotations might have an important communicative role in the quality control process.

Facilitation function

- Bramley and Pollitt (1996) observed that the majority of markers considered that annotating contributed to the improvement of their marking, helped them to apply performance criteria, and reduced the subjectivity of their judgements.
- O'Hara and Sellen (1997) suggest that readers of texts annotate in order to highlight structural features of the text and salient features, to record questions or draw attention to ideas that require reflection or further investigation.
- Annotations may offer cognitive support for comprehension building as well as performing other functions which are specifically linked to the context of the examination process (Anderson and Armbruster, 1982; Askwall, 1985; O'Hara, 1996; O'Hara and Sellen, 1997; Benson, 2001; Crisp and Johnson, 2005);
- According to Bramley and Pollitt (1996, p.6), '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'.
- In assessing feedback given to students when assignments were submitted and feedback returned on paper as well as on screen, Price and Petre (1997) observed that the quality and type of feedback were found to be similar. However, annotations providing *emphasis* were used less on-screen (although their use increased with increasing software familiarity).
- Shaw (2005) observed that examiners use annotations to investigate their own marking consistency. Annotations provide an efficient means to confirm, deny or reconsider standards both within and across candidates thereby reassuring examiners throughout the marking event.
- Crisp and Johnson (2005) investigated the use of annotations made by examiners marking a small number of GCSE Mathematics and Business Studies scripts. Their findings indicated that markers consider annotating to be a positive aspect of marking. This reflects the conclusions drawn by Bramley and Pollitt (1996) which suggest

that markers understand the process of annotations as being integral to, and contributing towards, the efficacy of marking.

Reading on-screen

- A growing body of research suggests that reading strategies employed to achieve comprehension of essays on paper play a vital role in the marking process and hence have implications for the reliability of marking (Sanderson, 2001; Crisp, 2007; Suto and Nádas, *in press*).
- Reading on-screen is 'generally less appealing than reading from paper' (Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt and Schedl, 2000, p.41).
- Research on first language (L1) reading indicates that reading rates drop 10–30% when moving from printed material to on-screen reading (Muter and Maurutto, 1991; Kurniawan and Zaphiris, 2001). Segalowitz, Poulsen and Komoda found that second language (L2) reading rates of highly bilingual readers are '30% or more slower than L1 reading rates' (1991, p.15).
- No single factor can account for why reading on-screen is perceived to be more difficult than reading on paper. In fact a number of variables are associated with reading on-screen: screen resolution, spatial representation, ease of use, disorientation, non-tangibility, experience, etc.
- Cassie (undated) cites two reasons why reading may be more difficult on a computer screen than on paper. First, readers tend to relate certain topics with strategically-situated locations on the page where they appear. Secondly, the process of reading through a number of printed pages is a tactile one: the reader having some comprehension of how far they have 'travelled' through a document.
- Related research has investigated the effects of computer familiarity on on-screen reading (Kirsch *et al*, 1998) and the effects of screen layout and navigation on reading from screen (Dyson and Kipping, 1998; dos Santos Lonsdale, Dyson and Reynolds, 2006).
- The visual layout of text and the mode of presentation affects the ease with which readers can access, read and respond to the text (Foltz, 1993; O'Hara and Sellen, 1997).
- Prior reading experience and computer familiarity are among factors that can influence reading assessment and methods (Rothkopf, 1978; Rayner and Pollatsek, 1989).
- Most empirical research into reading on-screen has separately addressed manipulation or navigation e.g. document structure, scrolling, page management (McDonald and Stevenson, 1996; Wenger and Payne, 1996; McDonald and Stevenson, 1998a, 1998b; Lin, 2003) and visual ergonomic factors e.g. layout variables (Dillon, 1994, 2004).
- One element of scrolling patterns (pauses between scrolling movements) has been identified as the main determinant of reading rate on-screen (Dyson and Haselgrove, 2000).

Context of the pilot

The Cambridge Checkpoint English examination is an innovative diagnostic testing service which provides standardised assessments for mid-secondary school pupils aged around 14. The tests, offered at two sessions each year, are designed to give feedback on individual strengths and weaknesses in the key curriculum areas of English, Mathematics and Science. The results provide teachers with information on student performance, enhanced by reporting tools built into the Checkpoint service.

English is assessed using two papers. Each paper takes one hour with an additional seven minutes for reading. In terms of the writing requirements, in Paper 1 candidates are given a short, focussed task with a clear aim and audience. The content is non-narrative and candidates are expected to write about 250 words. Paper 2 consists of a short and focussed task that does have a narrative content. Again, candidates are expected to write about 250 words.

Pilot design

The pilot employed a mixture of quantitative and qualitative methods. Quantitative methods used included correlational analyses of marks; computation of examiner inter-rater reliabilities; and Multi-Faceted Rasch Analyses (MFRA). The qualitative dimension of the pilot involved collating and analysing retrospective data captured by an examiner questionnaire. The research design, which was 'matched, between groups', tested the effect of two variables: marking medium and annotation sophistication, using four discrete marking conditions:

a) pilot scripts, **paper** marked, using **sophisticated** annotation
b) pilot scripts, **paper** marked, using **simplified** annotation
c) pilot scripts, marked **on-screen**, emulating current **sophisticated** annotation
d) pilot scripts, marked **on-screen**, using **simplified** annotation.

Table 1: Research Design

	Marking medium (Variable 1)		*Annotation (Variable 2)*	
	Paper	*On-screen*	*Sophisticated*	*Simple*
Method A	✔		✔	
Method B	✔			✔
Method C		✔	✔	
Method D		✔		✔

Ten examiners, including the Principal Examiner (PE), took part in the study, which consisted of two phases of marking. In phase 1, the examiners all marked the same set of 20 scripts on paper using sophisticated annotations. This 'calibration marking' provided a common baseline for the variation between these examiners under normal marking conditions. In phase 2, the examiners were split into four different sub-sets, one for each of the four marking conditions. All examiners then marked a further 200 scripts. Once again, the examiners marked the same scripts as each other (See Figure 1).

The examiners had various levels of experience but all had marked these question papers in the May 2007 administration and had been standardised then. The research was conducted in September 2007.

Marks and annotations from the live, on-paper May 2007 marking were removed from the 20 scripts which were subsequently coded,

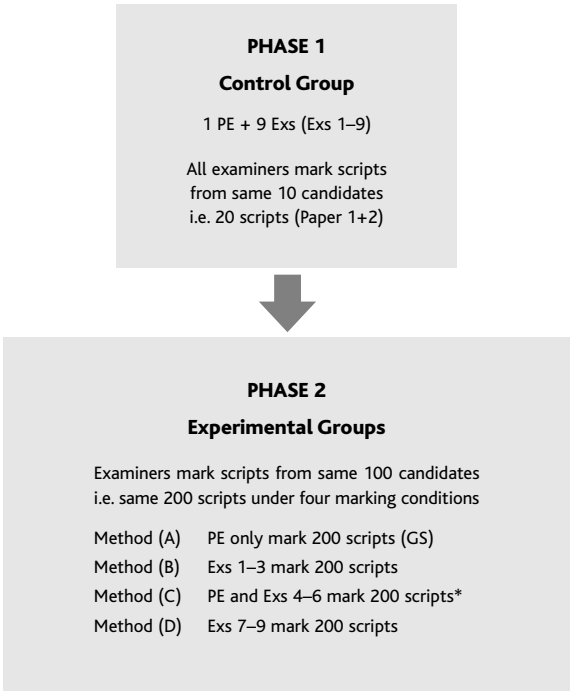

Figure 1: Research Design

copied and despatched to examiners for phase 1 of the pilot. The number of scripts required for the second phase of marking was arrived at through power test considerations (Kraemer and Thieman, 1987). Two hundred scripts (100 candidate performances) were scanned without annotations or marks to meet the requirements of marking under conditions described by Methods (C) and (D). In addition, unmarked hard copy versions were produced for Methods (A) and (B). Writing performances were identified as scripts which represented the full proficiency continuum for the test, exemplified a range of 'marked' profiles, and a diversity of centres.

In addition to empirical methodologies, emphasis was also attached to qualitative approaches. It was hoped that feedback from examiners would provide valuable insight into their on-screen marking experiences.

Findings

Phase 1: calibration markings

Descriptive statistics and analysis-of-variance indicated that the examiners were generally homogeneous in the marks they awarded to the 20 phase 1 scripts. Examiner inter-correlations were consistently high and indicated that examiners were reliably distinguishing between the respective assessment criteria on each paper. Strength of agreement tests revealed that whilst examiners were in general agreement on the rank ordering of the scripts, they were in less agreement regarding the absolute mark assigned to those scripts. However, inter-rater reliabilities were consistently high (of the order of 0.8), and Multi-Facet Rasch Analysis revealed that all examiners fell within the limits of acceptable model fit and that differences in severity / leniency between examiners were within tolerance (recommended cut off for flagging misfits includes t values outside +/- 2.0 [Smith, 1992]). The results of the phase 1 calibration markings therefore provide evidence that any quantitative differences found between the sub-groups in phase 2 are unlikely to be due to inherent differences between the markers in the sub-groups.

Phase 2: the four experimental marking methods

Before the marks from the four sub-groups were compared with each other, a quick comparison was made between the phase 1 and phase 2 marks. This indicated that examiners retained their relative levels of severity/leniency across both phases, that is, an examiner who was a little severe or lenient compared to the Principle Examiner in phase 1 was also a little severe or lenient in phase 2. As previously noted, however, there were no large differences in severity or leniency between examiners in phase 1.

Table 2 shows descriptive statistics across all four marking methods and for the live marks awarded in May 2007. The pilot means tended to be slightly higher than the live means.

The pilot standard deviations tended to be a little smaller than the live standard deviation for paper 1, but a little larger for paper 2. There were no large differences, however.

Table 3 shows the distribution of differences between the Principle Examiner marks for Method A (conventional marking) and the other examiners, aggregated by marking method. Method C (on-screen, sophisticated annotations) demonstrates the highest proportion of marks within +/- 3 marks of the PE.

Inter-examiner reliability indices were computed following the approach advocated by Hatch and Lazaraton (1991). A Pearson correlation matrix was generated for each marking method and then the average correlation for each method was calculated. A Fisher Z transformation was applied to the correlations before averaging to transform the correlations to a normal distribution suitable for averaging (Hatch and Lazaraton 1991). Table 4 presents the average correlations. The figures are high for both on-paper marking (method B) and on-screen marking (methods C and D). Although the inter-rater reliability is a little lower for the on-screen marking methods, the difference is not statistically significant.

Table 2: Overall comparison between Methods A – D and the live marks (Descriptive Statistics)

	Live May 2007			*Method A*			*Method B*			*Method C*			*Method D*		
	P1	*P2*	*Tot*	*P1*	*P2*	*Tot*	*P1*	*P2*	*Tot*	*P1*	*P2*	*Tot*	*P1*	*P2*	*Tot*
Mean	16.91	15.94	32.85	17.16	17.16	34.32	16.79	16.32	33.11	17.18	15.90	33.08	17.89	17.03	34.92
Std. dev.	6.71	6.00	12.10	6.12	6.14	11.69	6.54	5.96	11.49	6.28	6.20	11.81	5.57	5.94	10.70

Table 3: Agreement levels between the PE and other examiners

Marking Method	*Percentage of scripts:*			
	Exact agreement	*Within +/- 1 mark of PE*	*Within +/- 2 marks of PE*	*Within +/- 3 marks of PE*
Method B				
Paper 1	17	48	68	81
Paper 2	14	31	50	72
Method C				
Paper 1	21	52	71	82
Paper 2	13	32	47	80
Method D				
Paper 1	11	31	54	70
Paper 2	9	33	55	73

Table 4: Inter-examiner reliabilities

	Average correlation between examiners		
	Method B	*Method C*	*Method D*
Paper 1	0.80	0.78	0.75
Paper 2	0.80	0.78	0.78
Total (Paper 1 + Paper 2)	0.81	0.79	0.79

Findings from the retrospective questionnaire given to participants indicated that:

- Reading on-screen imposes higher cognitive demands on the marking process, particularly in relation to scrolling, page management, and application of annotations. Examiners suggested that protracted script electronic accessing procedures and slow script downloads may have deleterious consequences for the marking process. Pilot participants noted that their marking productivity was dependant upon several factors but chiefly the script downloading time.
- Examiners found scripts on-screen to be less easy to read than their paper counterparts (although this was not too great a problem for Checkpoint responses).
- Reading on-screen may adversely affect examiner concentration. Not being able to replicate paper and pen practice when applying annotations was a concern amongst pilot examiners. It was generally felt that on-screen marking is physically more demanding than paper marking and that marking over prolonged periods would engender mental and physical fatigue. For example, the physical process of selecting and applying pre-set annotations had implications for examiner concentration. It was believed that the additional cognitive demand intrudes upon the assessment process.
- Navigational demands imposed on the examiner by the computer interface affect the reading of text on-screen. Scrolling, for example, was considered by many examiners to be slow and generally annoying, presenting an unnecessary distraction to the reader.
- Script navigation was not as easy electronically as it is on paper. Reading on-screen inhibits formulation of a sense of overall meaning from the text and appears to impact negatively on examiner understanding of the marking criteria. Assessment criteria most affected tend to be those that define the macro features of text such as *rhetoric* (relating to discoursal features) and *organisation* (relating to coherence and cohesion).
- Whole text appreciation is impaired on-screen due to limited screen view and disrupted spatial layout. Holistic appreciation of the text was less achievable electronically as snapshots allow only restricted and incomplete sight of the text. This was especially noticeable when examiners were asked to consider the overall clarity and fluency of the message and how the response organises and links information, ideas and language.
- Reading on-screen may interfere with conventional, paper-based strategies employed to facilitate comprehension of the text message. The effect of mode seemed to encourage the use of different reading strategies, examiners having to revise their approach to assessment when marking on-screen.
- Prior experience with on-screen marking seems to have a positive influence on reading comprehension. Two of the pilot examiners, both of whom were consistent and reliable in their assessments (on paper and on-screen), claimed previous familiarity with on-screen marking.
- Identifying key features of textual information on-screen is more difficult than on paper.
- Reading on-screen may impede examiner construction of a mental representation of the text.
- Annotations aid textual comprehension. Whilst annotations are more awkward to apply on-screen, examiners were universal in their assertion that inability to annotate may impact negatively on the marking process. Participants were unanimous in their belief that the process of annotating enabled them to arrive at the right judgement(s).
- On-screen annotating may enhance marker reliability particularly as the software imposes a standardised set of electronic annotations.
- Examiners using the simplified form of annotation did not consider the range of annotation sufficient for marking purposes: the simplified suite of annotations being too restrictive.
- Examiners reinforced the prevailing belief that annotated scripts serve as a permanent record for subsequent adjudication and perform a communicative function between examiners.
- Generally, examiners were mixed regarding whether the time taken

to mark scripts on screen was the same as the time required to mark ordinary paper scripts. Despite difficulties encountered both reading and assessing on-screen, the majority of examiners believed that they ended up with about the same mark for each candidate across both modes. Whilst most examiners would still prefer to mark on paper, finding on-screen marking less enjoyable, nearly all examiners would be willing to use similar software in future sessions.

Discussion and Conclusion

The pilot found that paper-based and screen-based inter-examiner reliability is high for the Cambridge Checkpoint English Examination. Although inter-rater reliability is lower on-screen it is only marginally deflated. This finding accords with the findings of other, similar studies (e.g. Twing *et al*., 2003).

Levels of agreement were investigated between the Principle Examiner, marking on paper using sophisticated annotations, and other examiners marking on paper with simplified annotations, on-screen with sophisticated annotations, and on-screen with simplified annotations. The best agreement was found for those examiners marking on-screen with sophisticated annotations, implying that using sophisticated annotations is more important for marking accuracy than whether the marking is done on screen or on paper.

Analysis of mark agreement can only take us so far in an investigation of comparability, however, since a high degree of mark convergence might still mask issues to do with construct validity. This might be because the scripts used in the study did not cover the full range of relevant features, or because the examiners were not marking correctly in either mode.

Construct validity refers to the extent to which the testing instrument measures the 'right' underlying psychological traits or 'constructs'. Clearly, it is important to ensure that the constructs that tests are measuring are precisely those they intend to and that these are not contaminated by other irrelevant constructs or effects. If the mode of marking or the level of annotation permitted affect examiners' reading or understanding of the text, their assessments may be affected and construct validity compromised.

A reasonably well-developed conceptualisation of construct validity encompasses three dimensions of any testing activity – cognitive validity (the cognitive processing by the candidates activated by the test question), context-based validity (consideration of the social and cultural contexts in which the question is performed as well as the content parameters) and scoring validity which relates to all aspects of reliability (Shaw and Weir, 2007). If aspects of scoring validity are compromised by different modes of presentation then construct validity is potentially threatened. The questionnaire data collected in the present study revealed a number of functional differences between on-screen and on-paper marking modes, and between simple and sophisticated annotations, that might affect construct validity, and these would repay further investigation.

Future research

Future research should aim to:

- Establish the effects of navigation facilities and annotative tools on reading assessment, particularly in the context of longer stretches of text.
- Identify conditions under which examiner assessment is affected by interface design.
- Develop a greater knowledge of reading processes on-screen through:
  - identifying means by which differences in reading are mediated;
  - exploring whether reading can be enhanced by manipulating mediating factors.

CIE will undertake future pilots with these aims in mind. Reliability across marking mode will continue to be an important consideration. One study will entail marking the Singapore General Paper (GCE AO Level) Paper 2 on-screen. Paper 2 includes two questions that, in terms of expected text length, make greater demands on candidate resources than the Checkpoint English test. In general, the longer the text candidates have to produce, the greater the language, content knowledge, organisational and monitoring metacognitive abilities that might be required in processing. Concomitant with these demands on candidates is an increased cognitive load placed upon the assessor during marking.

References

Anderson, T. H. & Armbruster, B. B. (1982). Reader and text-studying strategies. In: W. Otto & S. White (Eds.), *Reading Expository Material*. London: Academic Press.

American Psychological Association (1985). *Guidelines for Computer-based Tests and Interpretations*. Washington, DC: Bennett.

Askwall, S. (1985). Computer supported reading vs. reading text on paper: a comparison of two reading situations. *International Journal of Man Machine Studies*, **22**, 425–439.

Bennett, R. E. (2003). *On-line Assessment and the Comparability of Score Meaning* (ETS RM-03-05). Princeton, NJ: Educational Testing Service.

Benson, P. J. (2001). Paper is still with us. *The Journal of Electronic Publishing*, **7**, 2. Available online at: www.press.umich.edu/jep/07-02/benson0702.html (accessed 25 August 2005).

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

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.

Cassie, T. (undated). *Reading and navigating of documents: digital versus paper*. Department of Computer Science, University of Maryland.

Crisp, V. (2007). Researching the judgement processes involved in A-level marking. *Research Matters: A Cambridge Assessment Publication*, **4**, 13–18.

Crisp, V. and Johnson, M. (2005). *The use of annotations in examination marking: opening a window into markers' minds*. Research Programmes Unit, Cambridge Assessment. A paper presented at the British Educational Research Association Annual Conference, University of Glamorgan, September 2005.

Dillon, A. (1994). *Designing usable electronic text: ergonomic aspects of human information usage*. London: Taylor and Francis.

Dillon, A. (2004). *Designing usable electronic text: ergonomic aspects of human information usage*. 2nd edition. London: CRC Press.

dos Santos Lonsdale, M., Dyson, M. C. & Reynolds, L. (2006). Reading in examination-type situations: the effects of text layout on performance. *Journal of Research in Reading*, **29**, 4, 433–453.

Dyson, M. C. & Haselgrove, M. (2000). The effects of reading speed and reading patterns on our understanding of text read from screen. *Journal of Research in Reading*, **23**, 1, 210–223.

Dyson, M. C. & Kipping, G. J. (1998). The effects of line length and method of movement on patterns of reading from screen. *Visible Language*, **32**, 2, 150–181.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 Reading Framework: A Working Paper*. TOEFL Monograph Series 17.

Foltz, P. W. (1993). *Readers' comprehension and strategies in linear text and hypertext*. Unpublished doctoral dissertation, University of Colorado, Boulder. Cited in Foltz, P. W. (1996), Comprehension, coherence and strategies in hypertext and linear text. In: E. Hatch & A. Lazaraton (1991), *The Research Manual: Design and Statistics for Applied Linguistics*. Boston, Massachusetts: Heinle & Heinle.

Johnson, M. & Greatorex, J. (2006). Judging learners' work on screen. How valid and fair are assessment judgements? *Research Matters: A Cambridge Assessment Publication*, **2**, 14–17.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees*. TOEFL Research Report 59. Princeton, NJ: Educational Testing Service.

Kraemer, H. C. & Thiemann, S. (1987). *How Many Subjects: Statistical Power Analysis in Research*. London: SAGE Publications.

Kurniawan, S. H. & Zaphiris, P. (2001). Reading online or on paper: Which is faster? In: *Proceedings of the 9th International Conference on Human Computer Interaction*, 220–222. August 5–10. New Orleans, LA.

Lin, D. (2003). Age differences in the performance of hypertext perusal as a function of text topology. *Behaviour and Information Technology*, **22**, 4, 219–226.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, **3**, 635–694.

Maughan, S. (2001). On-line Teacher Support: A Teachers' Perspective. CIE, UCLES internal report.

Messick, S. A. (1989). Validity. In: R. L. Linn (Ed.), *Educational Measurement*. 3rd edition. New York: Macmillan.

McDonald, A. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, **39**, 299–312.

McDonald, S. & Stevenson, R. J. (1996). Disorientation in hypertext: the effects of three text structures on navigation performance. *Applied Ergonomics*, **27**, 1, 61–68.

McDonald, S. & Stevenson, R. J. (1998a). Effects of text structure and prior knowledge of the learner on navigation in hypertext. *Human Factors*, **40**, 1, 18–27.

McDonald, S. & Stevenson, R. J. (1998b). Navigation in hyperspace: an evaluation of the effects of navigational tools and subject matter expertise on browsing and information. *Interacting with Computers*, **10**, 2, 129–142.

Murphy, R. (1979). Removing the marks from examination scripts before remarking them: does it make any difference? *British Journal of Educational Psychology*, **49**, 73–8.

Muter, P. & Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited? *Behaviour and Information Technology*, **10**, 257–266.

Newton, P. E. (1996). The reliability of marking of General Certificate of Secondary Education Scripts: mathematics and English. *British Educational Research Journal*, **22**, 4, 405–420.

O'Hara, K. (1996). *Towards a typology of reading goals: RXRC affordances of paper project*. Report EPC-1996-107. Available online at: www.lergonome.org/pdf/EPC 1996-107.pdf (accessed 25 August 2005).

O'Hara, K. & Sellen, A. (1997). *A comparison of reading paper and online documents*. *Proceedings of the Conference on human factors in computing systems (CHI '97)*. 335–342. New York: Association for Computing Machinery.

Price, B. & Petre, M. (1997). *Teaching Programming through Paperless Assignments: an empirical evaluation of instructor feedback*. *Proceedings of ITiCSE '97*. New York: ACM. Available online at: mcs.open.ac.uk/bp5/papers/1997-ITiCSE/ITiCSE%2097-price-petre.pdf (accessed 25 August 2005).

Raikes, N., Greatorex, J. & Shaw, S. (2004). *From paper to screen: some issues on the way*. Paper presented at the International Association of Educational Assessment conference, Philadelphia, June. Available at: www.ucles.org.uk/assessmentdirectorate/articles/confproceedingsetc/IAEA2000NRJGSS (accessed 25 August 2005).

Rayner, K. & Pollatsek, A. (1989). *The psychology of reading*. Hillsdale, New Jersey: Lawrence Erlbaum.

Rothkopf, E. Z. (1978). Analyzing eye movements to infer processing styles during learning from text. In: J. W. Senders, D. F. Fisher & R. A. Monty (Eds.), *Eye movements and the higher psychological functions*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Royal-Dawson, L. (2003). Electronic Marking with ETS Software. AQA Research Committee paper RC/219. In: D. Fowles & C. Adams (2005). *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the International Association for Educational Assessment Annual Conference, Abuja, retrieved February 5, 2006, from: www.iaea.info/abstract_files/paper_051218101528.doc

Salmon, G. (2004). *E-moderating. The key to teaching and learning on-line*. London, UK: Routledge Falmer.

Sanderson, P. J. (2001). *Language and differentiation in Examining at A Level*. PhD Thesis, University of Leeds, Leeds.

Segalowitz, G. M., Poulsen, C., & Komoda, M. (1991). Lower level components of reading skill in higher level bilinguals: Implications for reading instruction. *AILA Review*, **8**, 15–30.

Shaw, S. (2005). *On-screen marking: investigating the examiners' experience through verbal protocol analysis*. University of Cambridge ESOL Examinations, Report No 561.

Shaw, S. D., Levey, S. & Fenn, S. (2001). *Electronic Script Management: Report on an exercise held 20, 21, 22 April 2001*. Cambridge: UCLES internal report.

Shaw, S. D & Weir, C. J. (2007). *Examining Writing: Research and practice in assessing second language writing*. Studies in Language Testing No. 26. Cambridge: Cambridge University Press.

Smith, R. N (1992). *Applications of Rasch Measurement*. Chicago: MESA Press.

Smith, B. & Caputi, P. (2007). Cognitive interference model of computer anxiety: Implications for computer-based assessment. *Computers in Human Behavior*, **23**, 3, 1481–1498.

Sturman, L. & Kispal, A. (2003). *To e or not to e? A comparison of electronic marking and paper-based marking*. Paper presented at the 29th International Association for Educational Assessment Conference, 5–10 October 2003, Manchester, UK.

Suto, W.M.I. & Nádas, R. (*in press*). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*.

Twing, J. S., Nichols, P. D., & Harrison, I. (2003). *The comparability of paper-based and image-based marking of a high stakes, large scale writing assessment*. Paper presented at the 29th International Association for Educational Assessment Conference, 7 October 2003, Manchester, United Kingdom.

Wenger, M. J. & Payne, D. G. (1996). Comprehension and retention of nonlinear text: considerations of working memory and material-appropriate processing. *American Journal of Psychology*, **109**, 1, 93–130.

Whetton, C. & Newton, P. (2002). *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment Conference, 1–6 September 2002, Hong Kong SAR, China.

Williamson, P. (2003). Setting, marking and awarding: the examination process. In: K. Tattershall, J. Day, H. James, D. Gillan & A. Spencer (Eds.), *Setting the Standard*. Manchester: AQA.

Zhang, Y., Powers, D. E., Wright, W. & Morgan, R. (2003). *Applying the On-line Scoring Network (OSN) to Advanced Placement Program (AP) Tests. (RR-03-12)*. Princeton, NJ: Educational Testing Service.