

On-screen essay marking reliability: towards an understanding of marker assessment behaviour

Stuart Shaw and Helen Imam, University of Cambridge International Examinations

Abstract

Computer assisted assessment offers many benefits over traditional paper methods. However, in transferring from one medium to another, it is crucial to ascertain the extent to which the new medium may alter the nature of traditional assessment practice or affect marking reliability. Whilst there is a substantial body of research comparing marking and marker behaviour on screen and on paper, only a paucity of the available literature relates to the marking of extended responses. Research into on-screen assessment of continuous writing and its impact upon markers' judgements is, therefore, both timely and important.

This paper describes the beginning of a series of on-screen marking trials at Cambridge Assessment which, commencing with the Checkpoint English examination, attempt to investigate marker reliability of extended responses, construct validity and whether factors such as annotation and navigation differentially influence marker performance across marking modes. The findings described here seek to ascertain whether markers make qualitatively different assessments when marking the same piece of writing but through a different medium. The paper explains how the trial influenced the decision to move to on-screen marking of Checkpoint English and highlights the challenge of maximising ease of marking without compromising assessment validity.

Keywords: reliability; construct validity; on-screen marking; annotation

Introduction

There is now a growing body of research that considers how markers' judgements about candidates' performances might be affected by whether the marking is done on screen or on paper (Whetton and Newton, 2002; Bennett, 2003; Royal-Dawson, 2003; Sturman and Kispal, 2003; Zhang, Powers, Wright and Morgan, 2003; Raikes, Greatorex and Shaw, 2004). However, only a paucity of the available literature looks at whether marking mode affects markers' judgements about longer written answers. Research into on-screen assessment of extended writing is, therefore, both challenging and timely and it is clear that there is a need to broaden the experimental scope of current research activity in order to address some of the issues involved.

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 affect marking reliability. Therefore, appropriate validation studies must be conducted before a new marking approach can be implemented in a live context.

The pilot described here is the first attempt by the University of Cambridge International Examinations (CIE) to mark essays on screen. The Cambridge Checkpoint English examination was chosen for two reasons. First, it is a relatively low-stakes exam, being a diagnostic formative assessment tool and not a qualification. Secondly, it lends itself easily to on-screen marking as the candidate responds on the question paper itself, which is currently a prerequisite for on-screen marking.

The pilot attempts to investigate marker reliability and construct validity, and whether factors such as annotation and navigation differentially influence marker performance across the on-paper and on-screen marking modes. One obvious difficulty for examiners is reading handwritten text of varying quality on screen. An additional, anticipated difficulty relates to the application of the 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.

From the outset, public need for reliability versus personal need for pragmatics (examiner ease of marking) was perceived to be an overriding research consideration. It was necessary for CIE to determine whether the twin issues were in conflict and whether the on-screen marking load on examiners could be eased by reducing the number of annotations without compromising reliability of marking.

The trial influenced the decision to go live with on-screen marking and helped to determine the final annotations.

Context of the pilot: Checkpoint English

The focus of this study is the Cambridge Checkpoint English examination, which is taken in private, English-medium schools mainly overseas. An innovative diagnostic testing service, Cambridge Checkpoint provides standardised assessments for mid-secondary school pupils aged around 14, similar to the Key Stage 3 tests in the UK. It represents the end point of the CIE Lower Secondary curriculum and also the starting point of the IGCSE or O Level programme. The tests, offered twice yearly, are designed to give feedback on individual strengths and weaknesses in the key curriculum areas of English, Mathematics and Science. The results provide information on individual student performance, teaching groups, the whole school entry and the entire cohort. This information is further enhanced by the specialist reporting tools built into the Checkpoint service. Teachers receive the results within four weeks of the test being taken, and can use the information to adjust teaching strategies for both individual pupils and whole cohorts. The results also give an indication of the likely grades to be achieved by individual candidates at IGCSE or GCE O Level English Language.

English is assessed using two one-hour papers (with an additional 7 minutes for reading). Each paper has a reading and writing section, one paper having a functional/factual approach and the other a literary/narrative approach. The writing task on each paper is short and focused, and candidates are expected to write about 250 words per task.

A Curriculum Framework for Checkpoint English outlines the aims and the skill-based assessment objectives. The overall test provides feedback on four reading skills as well as eight writing skills: *structure; content; style; audience; sentence structure; vocabulary; punctuation;* and, *spelling*. Each paper tests six writing skills through six marking criteria, thereby reporting six sets of marks for each writing task. The annotations currently used to indicate errors include *spelling (sp); punctuation (|| for full-stop omission and p for other punctuation errors); style (wavy underline); word omission (^); and (√) to credit good expression*.

The Research Literature

There is a large research literature relevant to this pilot. Key aspects of the literature are summarised in relation to three principal themes: marking comparability; examiners' annotations; and on-screen reading.

Marking comparability

The literature appears to be 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" (Bennett, 2003:15). Notable exceptions where differences have been found not reviewed by Bennett include studies by Whetton and Newton (2002) and Royal-Dawson (2003).

Sturman and Kispal (2003) observed quantitative differences between on-screen 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. Sturman and Kispal concluded that screen-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.

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. Johnson and Greatorex (2008) 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 especially in relation to the marking of extended responses on screen.

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 to ascertain in exactly what circumstances on-screen marking is both valid and reliable.

Examiners' annotations

The limited extant empirical literature relating to the role of annotation in assessment practice has a number of parallels with the wider literature. The functions of annotating outlined by Wolfe and Neuwirth (2001) appear pertinent to assessment, particularly notions that annotations can facilitate assessors' reading, eavesdrop on the insights of assessors and call attention to topics in important text passages. Studying annotating practices in two different examiner groups, Crisp and Johnson (2007) found that annotating supported both individual and public functions. Interestingly, this public annotating function appeared to give individual examiners confidence in their continued marking practice since their visible annotations were a tool for transmitting the basis of their judgments to others in their community. The individual functions mirrored those outlined by Hsieh, Wood and Sellen (2006), with Crisp and Johnson reporting evidence that annotating helped with concurrent reading comprehension as well as post hoc judgmental processes.

Studies by Bramley and Pollitt (1996) and Shaw (2005) have also alluded to the way that annotations can interact with examiner confidence. Bramley and Pollitt found that examiners felt that annotating improved their marking, although this effect was not found in statistical analyses, whilst Shaw found that examiners liked to use annotations to check the accuracy of their own marking: annotations provide an efficient means to confirm, deny or reconsider standards both within and across candidates thereby reassuring examiners throughout the marking event. The Crisp and Johnson study indicated that markers consider annotating to be a positive aspect of marking (see also Johnson and Shaw, 2008). 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.

The relationship between annotator and subsequent readers of the annotation has also been explored in assessment contexts. Similar to the findings of Wolfe (2000), Murphy (1979) explored the influence of annotations beyond the individual annotator. He found significantly higher correlations between an original and a second mark where the mark and annotations were present on an examination scripts compared to where marks and annotations were removed.

Variations in assessor annotations have also been found to be influenced by several factors (Crisp and Johnson, 2007). Greatorex (2004) and Price and Petre (1997) found that mode influenced some annotation practices, with assessors using different annotation conventions on screen compared with paper. 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 emphasising features of a text were used less on screen (although their use increased with increasing software familiarity).

O'Hara and Sellen (1997) argue that mode can affect reader annotation in a number of ways. One major concern is the degree of physical effort required to annotate, say, in one mode compared with another. They suggest that applying paper-based annotations is a relatively effortless procedure and, as a consequence, it factors automatically into the meaning construction process during reading. In contrast, computer-based annotation practices can be impeded by the availability of annotation tools. Keyboards might influence annotating behaviour because they do not accommodate many of the types of annotating tools that readers choose to use when working on paper, therefore, making the process less authentic and positively affecting the cognitive demand on the reader.

Shaw (2005) observes that marker concentration can be adversely affected when assessing on screen. Not being able to replicate paper and pen practice when applying annotations was a predominant concern amongst the markers in his study. Echoing the findings of O'Hara and Sellen (1997), markers generally perceived on-screen annotating to be physically more demanding than paper annotating.

On-screen reading

Research suggests that readers experience a variety of difficulties when reading documents on screen:

- scrolling patterns: pauses between scrolling movements (Dyson and Haselgrove, 2001),
- awkward/cumbersome navigation (Dillon 1994; O'Hara and Sellen 1997)
- a lack of complete overview of document (O'Hara and Sellen 1997)
- lower tangibility of electronic documents compared to paper (Hansen and Haas 1988)
- an unclear awareness of the length of documents (O'Hara and Sellen 1997)
- lower reading speed caused by the poor screen resolution (Mills and Weldon 1987; Dillon 1994)
- learning of lower quality compared to paper documents (Hertzum and Frøkjær 1996)
- potential fatigue if reading is over an extended period

Much of the research indicates that reading on-screen is “generally less appealing than reading from paper” (Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt and Schedl, 2000:41). Research on first language (L1) reading indicates that reading rates drop 10-30% when moving from printed material to on-screen reading (Bailey, 1999; Muter and Maurutto, 1991; Kurniawan and Zaphiris, 2001). From a second language perspective, Segalowitz, Poulsen and Komoda found that reading rates of highly bilingual readers are “30% or more slower than L1 reading rates” (200:15).

No one single factor can account for why reading on screen is perceived to be more difficult than reading on paper. Muter and Maurutto (1991) have compiled a list of twenty-nine differences between reading from paper and screen that may account for the slower reading speeds on screen including: distance between the reading material and the reader; character shape; inter-line spacing; contrast ratio between characters and background; intermittent vs. continuous light (Wilkins, 1986); interference from reflections (Daniel & Reinking, 1987); stability - potential flicker, jitter, shimmer, or swim (Stewart, 1979); absence versus presence of incidental location cues (Wright and Lickorish, 1984); system response time.

Cassie (undated) cites two reasons why reading may be more difficult on a computer screen than on paper. Firstly, 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 the document.

Most empirical research into reading on screen has separately addressed manipulation or navigation e.g. document structure, scrolling, page management (pagination), annotating (McDonald and Stevenson, 1996; Wenger and Payne, 1996; O'Hara and Sellen, 1997; McDonald and Stevenson, 1998a, 1998b; Lin, 2003) and visual ergonomic factors e.g. layout variables (Dillon, 1994, 2004). The visual layout of text and the mode of presentation appear to affect ease of accessing, reading and responding (Foltz, 1993; O'Hara and Sellen, 1997).

Research has also been conducted on how readers accomplish their reading objectives on screen. Prior reading experience is a factor that can influence reading assessment and strategies (Rothkopf, 1978; Rayner and Pollatsek, 1989). Three groups of characterizing patterns in digital reading activity emerge from the literature: (1) non-linear reading occasionally with multiple readings of some sections (Bazerman, 1988: 235–253; Dillon, 1994: 93–101; Horney and Anderson-Inman, 1994); (2) linear reading under certain circumstances (Goldman and Saul, 1990; Foltz, 1996; Hertzum, Lalmas and Frøkjær, 2001); and, (3) strategic or ‘role’ reading. For example, Bazerman (1988) suggests that for academic papers, those sections containing dense formulas or problem formulations, might be omitted entirely.

Methodology

The marking 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 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 (method A)
- b) pilot scripts, **paper** marked, using **simplified** annotation (method B)
- c) pilot scripts, marked **on screen**, emulating current **sophisticated** annotation (method C)
- d) pilot scripts, marked **on screen**, using **simplified** annotation (method D).

Checkpoint English is currently marked in a manner that reflects method A. For methods A and C, the same set of six sophisticated annotations was used (see earlier description) and for methods B and D, the same set of three simple annotations was used (✓ for good expression, x for errors, and **wavy underline/highlight** for awkward style). The paper-based wavy underline was not available on screen at the time of the trial and a yellow highlight was used instead for methods C and D.

Examiners

Scripts from the May 2007 examination session were used for the trial. Ten examiners from the live session, including the Principal Examiner (PE), took part in the pilot, which consisted of two phases of marking. The research was conducted in September 2007.

The examiners had various levels of experience but all had marked these question papers in the May 2007 administration and so had already been standardised, and they were supplied with the same post-coordination mark schemes for the pilot. It was not felt necessary or desirable to have a further coordination meeting so soon after the live session because the potential small risk of forgetting what was conveyed orally at the coordination meeting was outweighed by a greater risk – of new issues ‘creeping’ into the additional meeting, a situation to be avoided as the whole point of the pilot was for comparisons to be made with the marks awarded for the live scripts.

Phases of research

In phase 1 of the pilot, all ten examiners marked the same set of 20 ‘calibration’ scripts (consisting of 10 candidate responses for Papers 1 and 2) on paper using sophisticated annotations. These scripts were selected on the basis that they exemplified a range of performance at both holistic and assessment criterion level. 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 with the PE marking under two groups - methods A and C (method A comprised the PE marks only). All examiners then marked a further 200 scripts (consisting of 100 candidate responses for Papers 1 and 2). The number of scripts required for phase 2 was arrived at through power test considerations (Kraemer and Thieman 1987). These scripts represented the full proficiency continuum for the test and exemplified a range of ‘marked’ profiles and a diversity of centres. Once again, the examiners marked the same scripts as each other.

Candidates had written their answers on the question papers in the normal way. For phase 1 of the pilot, marks and annotations were removed from the 20 scripts, which were subsequently coded, copied and despatched to examiners.

For phase 2, the 200 scripts were scanned without annotations or marks to meet the requirements of marking under conditions described by methods C and D. Digital images of the scanned scripts were sent by secure electronic link to examiners for on-screen marking at home using Scoris® software. In addition, unmarked hard copy versions were produced for methods A and B. Examiners were instructed to mark up essay responses once only even though the technology allowed separate mark-up for each assessment criterion.

As well as 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.

Quantitative Findings

Phase 1: calibration markings

Results of phase 1 calibration marking provided evidence of the validity of the rating scales and the reliability of marking. Descriptive statistics and analysis-of-variance indicated that the examiners were generally homogeneous in the marks they awarded to the twenty phase 1 ‘calibration’ 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 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 PE 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 1 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 statistically large differences, however.

Table 1: Overall comparison between Methods A – D and live marks

	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

P1 = Paper 1; P2 = Paper 2; Tot = Total

Marker agreement was a key concern throughout the pilot. Agreement was used here in two senses:

- as an indicator of PE agreement, that is, the PE's mark (representing the 'standard' for a particular performance) was used as the comparison mark against the examiners' marks (or average of marks);
- as an indicator of a set of marks (Bramley, 2007:26), that is, multiple observations of the same performances by a group of markers.

Table 2 shows the distribution of differences between the PE 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.

Table 2: 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

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: 533-535). Table 3 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 3: 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

Qualitative Findings

Questionnaire

Findings from the questionnaire indicated that:

- reading on screen imposes higher cognitive demands on the marking process particularly in relation to scrolling, page management and application of annotations. Markers suggested that protracted script electronic accessing procedures and slow script downloads may have deleterious consequences for the marking process. They also noted that their marking productivity was dependant upon several factors but chiefly the script downloading time;
- navigational demands imposed on the marker by the computer interface affect the reading of text on screen. Script navigation is not as easy electronically as it is on paper. Scrolling, for example, was considered by many markers to be slow and generally annoying, presenting an unnecessary distraction;
- markers also pointed to the fact that access to previous pages is relatively straightforward when reading from paper scripts. On screen, however, access rates are not as quick and the break between screens of text is, as markers commented, more critical. Whilst markers reported the 'nuisance' value of reading a sentence or paragraphs across screens, generally it would appear that splitting text across screens did not greatly affect marker comprehension levels;
- markers commented on the loss of context and meaning when reading some of the lengthier answers. Reading on screen inhibits formulation of a sense of overall meaning from the text and appears to impact negatively on marker understanding of the assessment criteria. The criteria most affected tended to be those that define the macro features of text such as *rhetoric* (relating to discursal features) and *organisation* (relating to coherence and cohesion);
- markers also articulated frustration at not being able to browse and scan read as effectively as when scanning printed pages;
- 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 markers were expected to consider textual features such as the overall clarity and fluency of the response and how the response organises and links information, ideas and language;
- prior experience with on-screen marking seems to have a positive influence on reading comprehension. Two of the markers, both of whom were consistent and reliable in their assessments across modes, claimed previous familiarity with on-screen marking;
- identifying key features of textual information on screen is more difficult than on paper. Markers supported the view that it was more difficult to identify, locate and re-locate key features of textual information on screen (particularly as the idea of a 'digital' page is a re-conceptualisation of a 'paper' page);
- reading on screen may adversely affect marker concentration. 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 marker concentration. Moreover, reading on screen may impede marker construction of a mental representation of the text;
- on-screen marking tends to engender multiple re-reading of candidate answers. Reading a script several times has obvious implications for both examiner throughput and concentration although markers did note that this becomes less of an issue as software familiarity improves;
- markers claimed that annotations aided their comprehension of a text. Whilst annotations are more awkward to apply on screen, markers were unanimous in their assertion that inability to annotate may impact negatively on the marking process. Markers also believed 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;
- markers using the simplified form of annotation did not consider the range of annotation to be sufficient for marking purposes: the simplified suite of annotations being too restrictive (see section on script scrutiny);
- markers reinforced the prevailing belief that annotated scripts serve as a permanent record for subsequent adjudication and perform a communicative function between examiners and as such annotation serves a key function in assessment;
- generally, markers were mixed regarding their view on 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 markers believed that they ended up with about the same mark for each candidate across both modes. Whilst most markers would still prefer to mark on paper, finding on-screen marking less enjoyable, nearly all markers would be willing to use similar software in future sessions.

Script scrutiny

In addition to consulting markers, it was also considered important to scrutinise random scripts marked by each method in order to gain an impression of the effects of the two variables, marking medium and annotation sophistication, on assessment practice. The scrutiny helped to confirm the conclusion of the quantitative research regarding the most appropriate method of annotation for future live on-screen marking as well as highlighting mark scheme and training issues. The main observations from the script scrutiny were that:

- *The marking medium has a mixed effect on the on-screen application of annotations.* The PE, who was the only marker placed in two groups (method A and C), was equally assiduous in the application of the mark scheme, displaying full use of the range of annotations, whether marking on screen or on paper. However, there were signs that individual markers could be affected by marking medium. For instance, one examiner gave up using annotations on screen altogether. On the other hand, on-screen marking often facilitated sampling: it was easier to see precisely where the annotations (such as crosses) had been placed on screen compared to on paper.
- *Sophisticated annotation appears more preferable.* Sophisticated annotation was more transparent and facilitated an immediate impression of the candidate's performance for each of the assessment criteria. This was deemed crucial during marking and for later sampling by senior examiners. Sophisticated annotation also appeared to lead to greater use of the lower mark range. Markers using the two sophisticated annotations **p** and **sp** to arrive at an assessment for more mechanistic criteria (punctuation and spelling) tended to be more discriminating. This could partly be explained because at a glance **p** and **sp** highlight the scale of error for each criteria, whereas the simple **x** annotation used to mark errors generally required sorting into the different criteria (spelling, punctuation, style, etc). In scrutinising scripts marked up with simple annotation, it took time to analyse the type of error and, therefore, the criteria that **x** referred to.
- *Personal differences in marking approaches raise mark scheme and training issues.* Certain criteria - *content* and *audience* - are holistic and it was difficult to see how markers arrived at marks as no annotations lent themselves to these criteria. Some markers were generous with the more subjective criteria (*content*, *style*, *sentence structure* and *vocabulary*) and used a greater number of ticks. These were surmised to be personal tendencies for a lower threshold for good expression, perhaps recognising that many candidates use English as a second language.

Script scrutiny led to a full review of the application of criteria, annotations and instructions for their use before live on-screen marking.

Discussion and Conclusion

The pilot revealed that paper-based and screen-based inter-marker reliability is high for the Cambridge Checkpoint English examination. The effect of marking mode is to depress the size of the reliability coefficient: inter-rater reliability is lower on screen although only marginally so. This finding accords with those of other, similar studies (e.g. Twing, Nichols and Harrison, 2003).

Levels of agreement were investigated between the Principal 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 marking methods). The best agreement was found for those examiners marking on screen with sophisticated annotations (method C), implying that sophisticated annotations are more important for marking accuracy than whether the marking is done on screen or on paper.

The statistical support for sophisticated annotation was confirmed by the pilot markers who did not consider the range of annotation provided by method D sufficient for marking purposes. Whilst markers acknowledged that annotation can aid comprehension, several commented on the negative impact of electronic annotation on reading stratagems:

"The business of moving the cursor up to a symbol in order to drag it down to the text, and then repeat the exercise umpteen times, is sapping, and a distraction from focusing on the text. It is not the way anyone reads a text..."

Method C may seem to be the obvious preferred method in terms of accuracy as it is the on-screen version of current practice. However, before the trial, method D (simple annotations) was anticipated to be the preferred method in terms of pragmatics. Markers were sceptical about applying the current full range of annotations on screen and were of the opinion that scripts would take too long to both read and mark. It was decided, therefore, to determine whether fewer annotations could be used (method D) without affecting reliability.

It was also thought that method D would ultimately suit the Checkpoint assessment as the reporting of multiple marks for each writing task could be managed well by the technology, which allows separate annotation for each of the marking criteria. In such a scenario, sophisticated annotation would be redundant as the nature of the error would be indicated by the screen displaying a particular marking criterion. It is possible that simplified annotation could produce the same level of accuracy if the text were marked up separately for each of the criteria. This would remove the mental difficulty of sorting crosses and ticks into the different marking criteria. However, this approach (multiple reading and mark-up) was not considered pragmatic as it does not reflect the way examiners behave (single reading and mark-up). It is feasible that the experience of extensive script marking in the future might lead examiners to consider the disadvantage of repetitive reading but with fewer annotations to be more attractive than the disadvantage of repetitive hand movement required for sophisticated annotation. If this were to happen, the medium of marking could eventually change examiners' 'natural' assessment behaviour.

Future research

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 pilot did not cover the full range of relevant features, or because the markers were not marking correctly in either mode. 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. Construct validity refers to the extent to which the testing instrument measures the 'right' underlying psychological traits or 'constructs'. A reasonably well-developed conceptualisation of construct validity encompasses three dimensions of the testing event - 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 including marker agreement (Shaw and Weir, 2007). If aspects of scoring validity are compromised by assessment mode then construct validity is potentially threatened.

Questionnaire data revealed a number of functional differences between screen and paper marking, and between simple and sophisticated annotations, that might affect construct validity, and these would repay further investigation particularly in relation to reading and assessment of even longer candidate answers on screen.

Several pilots have since been undertaken (both within CIE and across Cambridge Assessment) and more are planned, which aim to establish the effects of navigation facilities and annotative tools on reading assessment and reading practice particularly in the context of essays. The pilots identify conditions under which examiner assessment is affected by interface design and as such offer greater insight into how reading is accomplished on screen. These issues are especially important in relation to assessments that include questions where the expected length of the answer makes greater demands on candidate resources than the Checkpoint English test. In general, the longer the text candidates are expected to produce, the greater the language, content knowledge, organisational and monitoring metacognitive abilities that might be required in processing. Concomitant with increased candidate demands is an increased cognitive load placed upon the examiner in assessing the candidate's response.

Impact on assessment practice

It is believed that the findings from this pilot will facilitate a smooth transition to on-screen Checkpoint English assessment. This is because CIE is not only confident that it will be reliable but the research has led to improvements that have been fed into the operation of future on-screen marking. For example, the annotation used to highlight stylistic error has since been modified to be easier to apply and to reflect conventional practice.

The pilot has provided CIE with the opportunity to refine the set of sophisticated annotations before live on-screen marking. The production of a definitive set of annotations to facilitate fuller use of the mark range across the six criteria was achieved through iterative consultations with senior examiners. Whilst the number of annotations has remained the same, more efficient use has been made of them – for example, the numbered tick annotations developed for the summary exercise in the reading section will now also be used to credit good expression for specific numbered marking criteria in the writing tasks. As a result, examiners have a greater awareness of mechanistic and subjective criteria and should, as a consequence, develop a more consistent approach to both crediting and error-marking. It is hoped that these refinements will bring current examiners even more in line with the Principal Examiner.

Finally, examiners feel consulted and more reassured that research has been conducted. They have engaged with on-screen marking at an early stage, have already embarked upon the learning curve and are willing to try it when it becomes live, as illustrated by the following comments:

“Yes. I cursed, went away, and returned.”

“On the whole, I was very happy with the system, and you will be able to prove whether marking of this type is reliable or not.”

In the final analysis, the pilot and the move to on-screen marking have provided the opportunity to make a subjective assessment activity potentially more objective.

References

- Bailey, B. (1999). *UI Design Update Newsletter*, February, 1999. [On-Line] Available: <http://www.humanfactors.com/library/feb99.asp>
- Bazerman, C. (1988). *Shaping Written Knowledge: the Genre and Activity of the Experimental Article in Science*. Madison, WI: University of Wisconsin Press.
- Bennett, R. E. (2003). *On-line Assessment and the Comparability of Score Meaning (ETS RM-03-05)*, Princeton, NJ: Educational Testing Service.
- Bramley, T., (2007) *Quantifying marker agreement: terminology, statistics and issues*. Research Matters: A Cambridge Assessment Publication, Issue 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., & Johnson, M. (2007). The Use of Annotations in Examination Marking: Opening a Window into Markers' Minds. *British Educational Research Journal*, 33, 6, 943-961.
- Daniel, D. B., and Reinkin, D. (1987). *The construct of legibility in electronic reading environments*, In D., Reinking (Ed.), *Reading and Computers: Issues for Theory and Practice* (New York: Teachers College Press).
- Dillon, A. (2004). *Designing usable electronic text: ergonomic aspects of human information usage*. (2nd ed.). Boca raton: CRC Press.
- Dillon, A. (1994). *Designing usable electronic text: ergonomic aspects of human information usage*. London: Taylor and Francis.
- Dyson, M. C., and 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., and 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., and Schedl, M. (2000). TOEFL 2000 Reading Framework: A Working Paper. TOEFL Monograph Series 17.
- Foltz, P. W. (1996). *Comprehension, coherence and strategies in hypertext and linear text*. In *Hypertext and Cognition*. Rouet, J. F., Levonen, J. J., Dillon, Andrew P., and Spiro, R. J., (Eds.). New Jersey: Lawrence Erlbaum Associates. 109–136.
- 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*.
- Goldman, S. R., & Saul, E. U. (1990). Flexibility in text processing: A strategy competition model. *Learning and Individual Differences*, 2, 181-219.
- Greatorex, J. (2004). *Moderated E-portfolio Project Evaluation*. Cambridge: University of Cambridge Local Examinations Syndicate.
- Hansen, W. J., Andhaas, C. (1988). Reading and writing with computers: a framework for explaining differences in performance. *Comm. ACM*, 31, 9, 1080–1089.
- Hatch, E., & Lazaraton, A. (1991). *The Research Manual: Design and Statistics for Applied Linguistics*. Boston, Massachusetts: Heinle & Heinle.
- Hertzum, M., Lalmas, M., and Frøkjær, E. (2001). *How are searching and reading intertwined during retrieval from hierarchically structured documents?* In *Proceedings of the IFIP TC 13 International Conference on Human-Computer Interaction*, Tokyo, Japan, Jul. 2001, M. KUROSU, Ed., IOS Press, Amsterdam, 537–544.
- Hertzum, M., Frøkjær, E. (1996). *Browsing and querying in online documentation: A study of user interface and the interaction process*. *ACM Transaction on Computer-Human Interaction*, New York: ACM, 3 (2) 136-161.
- Horney, M. A., & Anderson-Inman, L. (1994). *Reading in hypertext: New skills for a new context*. *Proceedings of The Tenth International conference on Technology and Education*, Cambridge, Massachusetts.
- Hsieh, G., Wood, K. R. & Sellen, A. (2006). *Peripheral Display of Digital Handwritten Notes*. *Proceedings of the Conference on Human Factors in Computing Systems*, Montreal, Quebec, 2006.
- Johnson, J., and Greatorex, J. (2008). Judging Text Presented on Screen: implications for validity. *E-Learning*, 5,1, 40-50

- Johnson, M., and Shaw, S. D. (2008). Annotating to comprehend: a marginalised activity? Research Matters: A Cambridge Assessment Publication, Issue 6, 18-24.
- Kraemer, H. C., and Thiemann, S. (1987). How Many Subjects? : Statistical Power Analysis in Research. SAGE Publications: London.
- Kurniawan, S. H., and Zaphiris, P. (2001). *Reading online or on paper: Which is faster?* In Proceedings of the 9th International Conference on Human Computer Interaction, pp. 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.
- McDonald, S., and 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., and 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.
- McDonald, S., and Stevenson, R. J. (1996). Disorientation in hypertext: the effects of three text structures on navigation performance. *Applied Ergonomics*, 27 (1), 61-68.
- Mills, C. B., and Weldon, L. J. (1987). *Reading text from computer screens*. Centre for automation Research, Human-Computer Interaction Laboratory, University of Maryland, MD 20742.
- 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., and Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited? *Behaviour and Information Technology*, 10, 257-266.
- 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., and Petre, M. (1997). *Teaching Programming through Paperless Assignments: An Empirical Evaluation of Instructor Feedback*. Milton Keynes: Centre for Informatics Education Research, Open University.
- Raikes, N., Graetorex, J., and Shaw, S. D. (2004). A Report on the March 2004 OCR On-Screen marking Trial: The examiners' experience. Internal UCLES report.
- Rayner, K., and 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 and R. A. Monty (Eds.), *Eye movements and the higher psychological functions*, pp. 209-223. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Royal-Dawson, L. (2003). *Electronic Marking with ETS Software*. AQA Research Committee paper RC/219, in Fowles, D., and Adams, C. (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.
- Segalowitz, G. M., Poulsen, C., and 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*. Internal ESOL Validation and Research Report.
- Shaw, S. D., Levey, S., and Fenn, S. (2001). *Electronic Script Management: Report on an exercise held 20, 21, 22 April 2001*, Cambridge: UCLES internal report.
- Shaw, S. D., and Weir, C. J. (2007). *Examining Writing: Research and Practice in assessing second language writing*. 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.
- Stewart, T.F.M. (1979). Eyestrain and visual display units: a review, *Displays*, 25-32.
- 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 29th International Association for Educational Assessment Conference, 5-10 October 2003, Manchester, UK.
- Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks*. Research Reports 61. Princeton, NJ: Educational Testing Service.
- 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.
- Waller, P. (1987). *The typographic contribution to language: towards a model of typographic genres and their underlying structures*. PhD Thesis, Dept. of Typography and Graphic Communication, University of Reading.

- Wenger, M. J., and 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., and 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.
- Wilkins, A. (1986). Intermittent illumination from visual display units and fluorescent lighting affects movements of the eyes across text, *Human Factors*, 28, 75-81.
- Wolfe, J. L. (2000). Effects of annotation on student readers and writers. JCCL '00, San Antonio, Texas, 19-26.
- Wolfe, J. L., & Neuwirth, C. M. (2001). From the Margins to the Center: The Future of Annotation. *Journal of Business and Technical Communication*, 15, 3, 333-371.
- Wright, P., and Lickorish, A. (1984). Investigating referees' requirements in an electronic medium, *Visible Language*, XVIII, 186-205.
- Zhang, Y., Powers, D. E., Wright, W. and Morgan, R. (2003) Applying the Online Scoring Network (OSN) to Advanced Placement Program (AP) Tests. (RR-03-12) Princeton, NJ: Educational Testing Service.