An investigation into marker reliability and some qualitative aspects of on-screen essay marking

Martin Johnson and Rita Nádas Research Division

A more detailed analysis of the reliability findings reported here will appear in 'Marking essays on screen: an investigation into the reliability of marketing extended subjective texts' to be published in the British Journal of Educational Technology *by the British Educational Communications and Technology Agency and Blackwell Publishing.*

Introduction

There is a growing body of research literature that considers how the mode of assessment, either computer- or paper-based, might affect candidates' performances (Paek, 2005). Despite this, there is a fairly narrow literature that shifts the focus of attention to those making assessment judgements and which considers issues of assessor consistency when dealing with extended textual answers in different modes.

This article argues that multidisciplinary links with research from domains such as ergonomics, the psychology of reading, human factors and human-computer interaction could be fruitful for assessment research. Some of the literature suggests that the mode in which longer texts are read might be expected to influence the way that readers access and comprehend such texts (Dillon, 1994; Hansen and Haas, 1988; Kurniawan and Zaphiris, 2001; Mills and Weldon, 1987; O'Hara and Sellen, 1997; Piolat, Roussey and Thunin, 1997; Wästlund, Reinikka, Norlander and Archer, 2005). This might be important since these factors would also be expected to influence assessors' text comprehension whilst judging extended textual responses.

Literature review

Gathering reliability measures is a significant practical step towards demonstrating the validity of computer-based testing during the transitional phase where assessments exist in both paper- and computer-based modes. In her review of comparability studies Paek (2005) notes that the transition from paper- to computer-based testing cannot be taken for granted and that comparability between the two testing modes needs to be established through carefully designed empirical work.

Paek suggests that one of the primary issues for such research is whether the computer introduces something unintended into the test-taking situation. In the context of assessing essays on screen this might demand enquiry into construct validity; exploring whether the same qualitative features of essay performance are being attended to by assessors in different modes.

Whilst Paek reports evidence that screen and paper versions of traditional multiple-choice tests are generally comparable across grades and academic subjects, she notes in her conclusion that 'tests with extended reading passages remain more difficult on computer than on paper' (p.18), and suggests that such differences might relate to computers inhibiting students' reading comprehension strategies.

Johnson and Greatorex (2008) extend this focus on comprehension to call for studies which explore the cognitive aspects of how judgements might be influenced when assessors read longer texts on screen. This concern appears to be important given a recent study which reports correlations between re-marked essays significantly lower when scripts are re-marked on screen compared with paper re-marking (Fowles, 2008).

There are a variety of cognitive aspects of reading whilst assessing. Just and Carpenter (1987) argue that working memory is directly linked to reading a text and that this involves an expectancy effect that relies on working memory to retain the words just read in order to allow the next words to be linked together in a meaningful way. They go on to suggest that increasing the complexity of the task or the number of component elements of the reading activity can also affect reading performance. Mayes, Sims and Koonce (2001) reiterate this point, reporting a study which found that increased reader workload related significantly to their reduced comprehension scores.

Another cognitive aspect of reading relates to the role of spatial encoding. Johnson-Laird (1983) suggests that the linear nature of the reading process leads to the gradual construction of a mental representation of a text in the head of the reader. This mental representation also accommodates the location of textual information with readers spatially encoding text during the reading process (Piolat, Roussey and Thunin, 1997). Spatial encoding hypothesis claims that positional information is processed during reading activity; the hypothesis is based on evidence that readers can regress to find a location within a visual text very efficiently.

Research suggests that the cognitive effort of reading can be augmented by other activities such as annotating and note taking, with these 'active reading' practices often operating concurrently with reading activity (O'Hara and Sellen, 1997; Piolat, Olive and Kellogg, 2005). Literature suggests that active reading can enhance reading comprehension by supporting working memory (Crisp and Johnson, 2007; Hsieh, Wood and Sellen, 2006; Marshall, 1997) and facilitate critical thinking (Schilit, Golovchinsky and Price, 1998). Schilit *et al*. (1998) observe that active reading is challenged by the screen environment due to difficulties in free-form ink annotation, landscape page orientation (leading to the loss of a full page view), and reduced tangibility.

Recent shifts in Human Factors research have been increasingly concerned with the cognitive demands related to reading across modes. Much of this work has focussed on the inherent features of computer displays and navigation issues. Since it has been found that protracted essay reading (and by inference essay assessment) can involve navigating a text in both linear and non-linear ways (O'Hara, 1996; Hornbæk and Frøkjær, 2001), on-screen navigation might exert an additional cognitive load on the reader. This is important given the suggestion that increased reading task complexity can adversely affect reading comprehension processes.

The literature has led to a model of the interactions that might influence mental workload whilst reading to comprehend. In the model, physical process factors such as navigation and active reading strategies are thought to support assessors' cognitive processing (e.g. spatial encoding) which could in turn affect their comprehension whilst they judge extended texts. Theory suggests that readers employ these physical processes differently according to mode and that this can affect reader comprehension. Studying physical reading processes might therefore help to explain any divergent assessment outcomes across modes. The model suggests that research might usefully include a number of quantitative and qualitative factors. Assessors' marking reliability across modes, their attention to different constructs, and their cognitive workloads could be quantitative areas of focus. These findings could be supplemented with qualitative data about factors such as navigation and annotation behaviours in order to explore influences on assessors' spatial encoding processes whilst comprehension building.

Research questions and methodology

The plan for this project considered 6 questions:

1. Does mode affect marker reliability?
2. Construct validity – do examiners consider different features of the essays when marking in different modes?
3. Is mental workload greater for marking on screen?
4. Is spatial encoding influenced by mode?
5. Is navigation influenced by mode?
6. Is 'active reading' influenced by mode?

One hundred and eighty GCSE English Literature examination essays were selected and divided into two matched samples. Each stratified sample contained 90 scripts spread as evenly as possible across the seven bands of the 30-point mark scheme.

The scripts were then blind marked for a second time by the subject Principal Examiner (PE) and Assistant Principal Examiner (APE) to establish a reference mark for each script. In this project the reference mark is therefore defined as the consensual paper mark awarded by the PE and the APE for each answer.

Twelve examiners were recruited for the study from those who marked the unit 'live' in January 2008. Examiner selection was based on the high quality of their past marking. In order to control the order of sample marking and marking mode, the examiners were allocated to one of four marking groups. Examiner groups 1 and 4 marked Sample 1 on paper and Sample 2 on screen; groups 2 and 3 marked Sample 1 on screen and Sample 2 on paper. Groups 1 and 3 marked Sample 1 first, and groups 1 and 2 marked on paper first. This design allowed subsequent analyses to separate out any purely mode related marking effects (i.e. direct comparisons of the marking outcomes of groups 1 and 4 with groups 2 and 3) from any marking order effects.

In order to replicate the normal marking experience as much as possible the examiners completed their marking at home. Before starting their on-screen marking all examiners attended a group training session to acquaint them with the marking software along with administrative instructions.

Marker reliability was investigated first by looking at the mean marks for each examiner in each mode. Overall comparisons of the mark distribution by mode and against the reference marks were also made. Statistical models were then used to investigate the interaction between each examiner and mode.

To investigate construct validity, the textual features that were perceived to characterise the qualities of each essay response were elicited through the use of a Kelly's Repertory Grid (KRG) exercise (Kelly, 1955; Jankowicz, 2004). This process involved the Principal Examiner (PE) and the Assistant Principal Examiner (APE) separately comparing essays that were judged to be worth different marks, resulting in 21 elicited constructs. The PE and APE then separately rated 106 scripts according to each individual construct on a 5-point scale. These construct ratings were added into the statistical models to investigate whether each construct influenced marking reliability in either or both modes.

To investigate mental workload in both marking modes, a subjective measure of cognitive workload was gathered for each examiner. The National Aeronautics and Space Administration Task Load Index (NASA-TLX) (Hart and Staveland, 1988) is one of the most commonly used multidimensional scales (Stanton *et al*., 2005). It is considered to be a robust measure of subjective workload (Moroney *et al*., 1995); demonstrating comparatively high factor validity; usability; workload representation (Hill *et al*., 1992); and test-retest reliability (Battiste and Bortolussi, 1988). This has led it to be used in a variety of studies comparing mode-related cognitive workload (e.g. Emerson and MacKay, 2006; Mayes *et al*., 2001).

For this study the NASA-TLX measure of mental workload was completed twice by each examiner, midway through their marking sessions in both modes. This enabled a statistical comparison of each marker across modes to explore whether screen marking was more demanding than paper marking.

The influence of mode on examiners' spatial encoding was investigated through their completion of a content memory task. After marking a randomly selected script in both modes, five of the examiners were asked to recall the page and the location within the page where they had made their first two annotations. A measure of spatial recall accuracy was constructed and used as a basis for comparison across modes.

To investigate how navigation was influenced by mode, information about reading navigation flow was gathered through observations of six examiners marking in both modes. This involved recording the directional flow of examiners' navigating behaviour as they worked through eight scripts.

Examiners' annotation behaviour was collected to explore how this aspect of 'active reading' was influenced by mode. Examiners' annotation behaviours were analysed through coding the annotations used on 30 paper and screen scripts from each of the examiners. This analysis of 720 scripts represented one-third of all the scripts marked.

Finally, concurrent information was gathered by the examiners in the form of an informal diary where they could note any issues that arose during marking. Alongside the marking observation data, this diary evidence provided a framework for a set of semi-structured interviews that were conducted with each examiner after the marking period had finished. This allowed the researchers to probe and check their understanding of the data.

Findings

Does mode affect marker reliability?

Initial analyses showed that neither mode order nor sample order had significant effects on examiners' reliability. Analyses of examiners' mean marks and standard deviations in both modes suggested no evidence of

any substantive mode-related differences (paper mean mark: 21.62 [*s.d. 3.89*]; screen mean mark: 21.73 [*s.d. 3.91*]). Five examiners tended to award higher marks on paper and seven awarded higher marks on screen. However, such analyses might mask the true level of examiner marking variation because they do not take into account the mark-disagreements between examiners at the individual script level.

To allow for this, further analysis considered the differences between examiners' marks and the reference marks awarded for the scripts. For the purposes of this analysis the chosen dependent variable was the script-level difference between the examiners' mark and the reference mark, known as the Mean Actual Difference, with negative values indicating that an examiner was severe and a positive value indicating that an examiner was lenient in relation to the reference mark.

Box plots for the distribution of the mark difference for scripts marked in both modes suggest little mode-related difference (Figure 1). These indicate that about half of the examiners showed a two–mark difference from the reference marks in both modes, with paper marking tending to be slightly more 'accurate'.

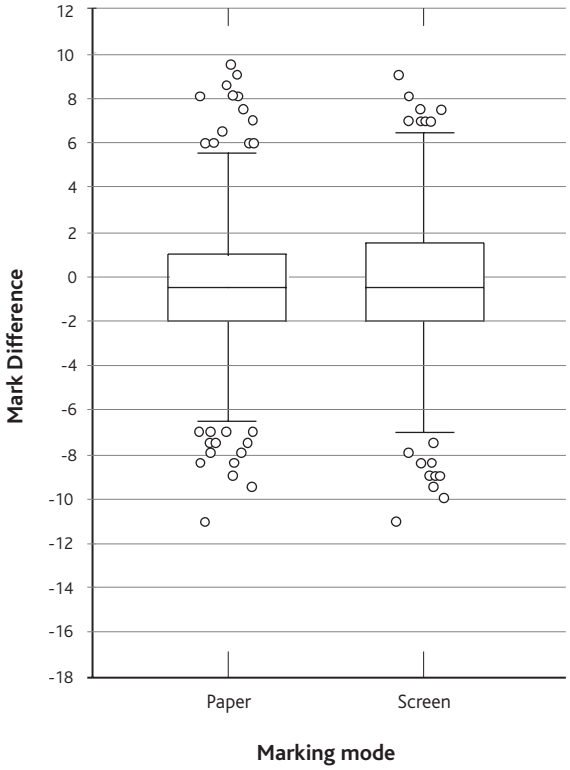

Figure 1: Box plots of the distribution of mark difference from the reference mark by marking mode[1]

To investigate the interaction between individual examiners and mode, least square means from an ANCOVA are plotted in Figure 2.

Figure 2 shows that the confidence intervals overlap for all examiners except for Examiner 4, suggesting no significant mode-related marking difference for 11 examiners. Where an examiner was severe or lenient in one mode they were also similarly severe or lenient in the other mode. Examiner 4 differed from the other examiners because his screen marking differed significantly from his paper marking with the screen marking being closer to the reference marks.

1 For ease of interpretation, the box includes 50% of the data, and each whisker represents 25% of the data. The horizontal line within the box is the median, below and above which lie 50% of the data.

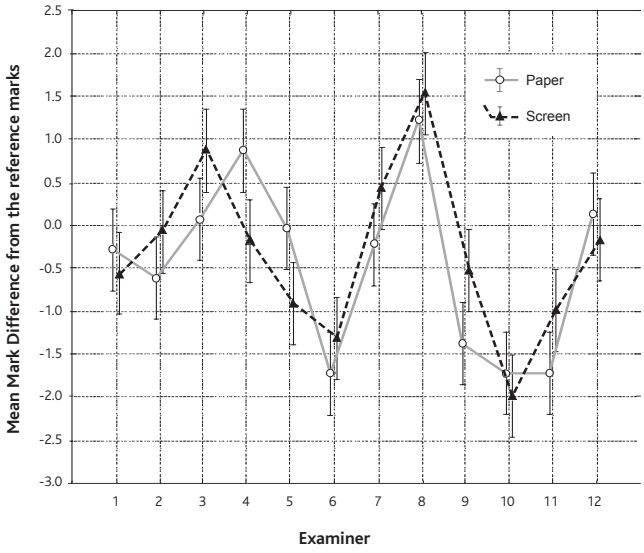

Figure 2: Least Square means for mark difference by examiner and mode

Construct validity – do examiners consider different features of the essays when marking in different modes?

21 sets of construct ratings were added in turn into the statistical reliability models in order to investigate whether each construct influenced marking in either or both modes. Data revealed that mode did not have a significant effect on the constructs examiners paid attention to while marking. However, some constructs did explain the difference between some individual examiners' marks and the reference marks; for example, 'points developed precisely and consistently'; 'insight into characters' motivation and interaction' or 'attention to both strands of the question'. Further research is currently underway on the relationship between examiners' use of constructs and essay marking performance.

Is mental workload greater for marking on screen?

Data suggest that overall cognitive load was greater for screen than paper marking ($t(11) = -2.95, p < 0.05$). Figure 3 shows the variations in the extent to which the subscales differed according to mode. The ***frustration*** subscale showed a large and statistically significant mode-related influence ($t(11) = -3.69, p < 0.01$), suggesting a greater factor in on-screen marking. A slight tendency was also found on the ***performance*** subscale ($t(11)=2.19, p=0.051$), suggesting that examiners were comparatively more satisfied with their marking on paper than on screen.

On all other dimensions marking mode did not have a significant effect on the cognitive load of the task, suggesting that the ***frustration*** experienced during screen marking contributed to the examiners' elevated overall cognitive load ratings for this marking mode.

Although overall cognitive workload ratings were higher for screen marking, significant variations were found between some of the total cognitive load ratings reported by examiners ($t(11) = 28.37, p < 0.001$). Furthermore, although all examiners reported concern for ***performance*** dimensions in both modes, there was variation among examiners concerning the rest of the dimensions.

In order to tease out which aspects of marking contributed to the above findings, and to explain the wide variation found in participants' profiles, follow-up semi-structured interviews were conducted with all participants.

On-screen marking was associated with significantly more ***frustration*** than traditional paper-based marking. Most examiners mentioned the

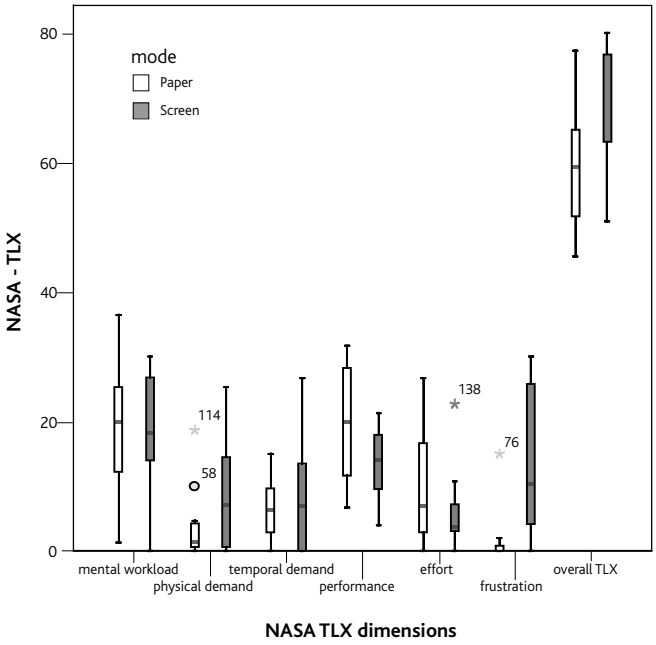

Figure 3: Mode-related differences in examiners' ratings on NASA-TLX dimensions and overall cognitive load

novelty of on-screen marking or specific elements of the software environment as causes for their initial frustration. However, once technical problems were resolved, examiners generally grew more comfortable with on-screen marking and frustration levels decreased.

Examiners were slightly less satisfied with their on-screen marking ***performance***. Some of the reasons for this related to the novelty of technology; the lack of a standardisation session; examiners' own personality traits, and the inherent responsibility of the marking process. Generally, it seemed that examiners perceived two types of performance: the satisfaction of completion and the professional accomplishment of performing high quality work.

Most of the sources of ***mental workload*** reported (e.g. cognitive processes, responsibility, unfamiliarity with the process, etc) are inherent characteristics of any marking process, and perhaps explain why mode did not have a significant effect on this subscale. Although causing a heightened initial mental workload, unfamiliarity with the situation eased as markers got used to the technology.

Mode had no significant effect on the ***physical demand*** of marking. A variety of activities, as well as the unfamiliarity and constraints of the physical environment, contributed to physical strain, which originated from inadequate working conditions characteristic of both marking modes.

Temporal demand was not significantly affected by marking mode, and was generally reported to be very low in the project overall. However, a live marking session with tight deadlines might result in heightened amounts of long-term temporal demand on examiners.

Data showed only a slight statistical tendency for on-screen marking to require more ***effort***. Participants listed a variety of elements which contributed to fatigue, for example, novelty and initial struggles with technology; sticking to and applying standards; physical strain and looming deadlines; mental fatigue; and administrative tasks/recording marks on paper. Others felt energised by some particular aspects of on-screen marking, for example, the ability to read poor handwriting; the lack of administrative requirements; and 'seeing the scripts off by a click'.

Is spatial encoding influenced by mode?

Whilst marking a randomly selected script in both modes, five of the examiners were asked to recall the page and the location within the page where they had made their first two annotations.

Although the number of examiners involved in this activity was limited it appears that the ability to recall not only the page but the location of a detail within that page was more precise on paper than on screen. On paper all five examiners could recall the page on which they made at least one of their first two annotations. Three of these annotations were located in the correct geographical ninth of the page and two were within the correct geographical third of the page. On screen only two of the examiners were able to locate the page of any of their annotations, and these were only positioned in the correct third of a page. The three remaining examiners could not remember the page where they made either of their first two annotations.

This suggests that the examiners' spatial encoding was better on paper and that this led to a better mental representation of the text read; as one examiner put it:

> *I do tend to have that sort of memory where I… know that it's at the top, middle or bottom of the page that I saw something. That sort of short term stays there, but with the zooming and scrolling it isn't quite as easy because on the paper you just turned, there it is and you've found it.* (Examiner: 10: Interview)

Theory suggests that readers spatially encode the location of features in a text when they construct a mental representation of it. It appears that the use of iterative navigational strategies can facilitate this process by affording readers the opportunity to efficiently locate and remember qualities within a text. At least two factors might influence this navigating activity: (i) reader annotation activity, and (ii) the characteristics of visual field and resolution levels in the reading environment.

Observations suggest that visual reading fields tend to be larger on paper and offer higher resolution levels, which in turn might influence navigation behaviour. Indeed, a number of examiners indicated that their marking practice involved them getting an overview of the script, reinforcing their mental image of it.

Is navigation influenced by mode?

Paired samples t-tests showed that examiners' paper navigation tended to be more iterative, using both linear and non-linear reading approaches, whilst on-screen navigation tended to be overwhelmingly linear ($t(5) = 2.84$, $p = 0.04$).

Iterative reading behaviours appeared to involve examiners establishing an overview of the script and it seems that the ability to gain an overview of the script positively influenced examiner confidence. Three examiners suggested that having an overview of the script made them feel more confident in the consistency of their marking. The reason for this perception seems to relate to the way that looking back over a script allowed examiners to confirm or question their previous reflections.

Three of the examiners suggested that navigational ease in the paper mode helped to support their working memory whilst building a sense of textual meaning. Another key mode-related factor appeared to be that the paper environment afforded fluid annotation across multiple pages. It appeared that not being able to navigate as freely around a script on screen led to some examiner frustration and their adoption of consciously different reading styles.

Interviews, observations and examiner diary evidence suggested that navigation away from the script was also related to mode. Examiners commonly described a reduced tendency to move their attention between different scripts on screen. Comparing the qualities of different scripts appears to be a key feature of the examiners' usual practice, with cross-referencing between scripts helping them to compare the qualities of different performances, establish or confirm a standard, and reinforce their confidence in the consistency of their own judgements.

It was very common for examiners to suggest that comparing the qualities of different scripts was less effective on screen. It is possible that such mode-related difference relates to how the tangibility of a text might support examiners' mental workload. One of the key links between tangibility and thinking might be the way that the paper environment can afford speedy comparisons to be made. One examiner noted that the process of identifying and accessing other potentially relevant scripts for comparative purposes is a rapid activity supported by speedy and targeted navigation:

> *When marking on paper, it's easy enough to look back at an earlier script. It's in a pile to one side and even if one does not remember the mark given, or the candidate's name or number, looking at the first sentence, paragraph, identifies the script wanted. With computer marking, 'flicking through the pile' is neither quick nor easy.* (Examiner 11: Diary)

Is annotation influenced by mode?

In order to compare examiners' annotation behaviours, 30 paper and screen scripts from each examiner were analysed and their annotation use coded. This analysis of 720 scripts represented one-third of all the scripts marked.

Examiners were able to use a wider variety of annotations on paper than on screen since the screen environment allowed only 10 annotation types. These annotations were built into the marking software following consultation with the examination's Principal Examiner.

Analysis showed that examiners used a wider variety of annotation types on paper (on average 7.58 annotation types per examiner) compared with on screen (6.75 annotation types per examiner). Written comments on paper accounted for most of the difference between the types of annotations used on screen and on paper. This type of annotation was used on average nearly 4 times per paper script and generally included sets of phrases directly linked to evidence found in the text to bring together subtle reflections (e.g. "possibly"), holistic and/or tentative judgements (e.g. "could be clearer"; "this page rather better"), to represent internal dialogue or dialogue with the candidate (e.g. "why?"), or to make note of particular features or qualities found in the text (e.g. "context"; "clear").

When comparing the use of the same ten annotations across modes, 8 of the 12 examiners annotated more on paper. Also, the mean number of annotations made on each paper script (19.48) was higher than on each screen script (18.62), although ANOVA analysis showed that this was not a statistically significant difference ($F(1, 22) = 0.13$, $p = 0.72$).

Despite this, ANOVA analyses showed significant mode-related differences between the mean number of paper and screen annotations for four specific annotation categories. "Underlining" ($F(1, 22) = 7.87$, $p = 0.01$) was used more heavily on paper whilst "Very Good" ($F(1, 22) = 4.78$, $p = 0.04$), "Excellent" ($F(1, 22) = 4.68$, $p = 0.04$) and "Support" ($F(1, 22) = 5.28$, $p = 0.03$) annotations were used significantly more frequently on screen. T-test analyses showed that examiners were also significantly more likely to use ideographic annotations to link text on paper such as circling and sidelining ($t(5) = 2.66$, $p < 0.05$), whereas screen annotations only allowed examiners to label discrete qualities found in the text.

It was usual for the examiners to write a final summative comment on the scripts in both modes. Analysis showed that summative comments were made on more than 99% of the paper script sample and more than 97% of the screen script sample. The importance of the summative comment was highlighted by two examiners who suggested that it factored into their final judgement about the quality of each script:

> *What I couldn't write in the margin, because the system didn't let me, I wanted to store up for the final comment. It seems to me that because you can't annotate, the final comment is more important on screen than it is on paper.* (Examiner 5: Interview)

> *In both cases it's in composing the comment that I harden up on exactly what mark I'm going to award.* (Examiner 8: Interview)

Discussion

It is important to acknowledge that this research project had a number of limitations relating to examiner sample, marking load and script distribution that could challenge the generalisability of the findings. First, the study involved only 12 examiners who were pre-selected for participation based on their high performance profiles, and thus their behaviour might not be representative of all examiners. Secondly, the examiners had a comparatively light marking load with a generous time allowance compared with live marking. Finally, the balance of the script sample characteristics did not necessarily reflect the balance of qualities that examiners might face during a live marking session.

This study was motivated by concerns that screen marking might interfere with examiners' reading processes and lead to marking variances when examiners assess longer texts on screen and on paper. This study found the variance across modes to be non-significant for all but one examiner, suggesting that the marking of these essays is feasible using this particular screen technology. Whilst this in itself is an interesting finding, it is only partial since the real issue concerns construct validity and whether marks were given for the same essay features in the different modes. Again, the Kelly's Repertory Grid analysis suggested that there were no significant relationships between specific essay constructs and differences between examiners' marks across modes. Most interestingly, some of these elicited constructs did explain the differences found between the marks given by different examiners regardless of mode, allowing an insight into the variances that are sometimes found between different examiners and providing obvious scope for further research.

Considering the research literature, these quantitative findings appear to sit uncomfortably with the qualitative study findings, and this requires some degree of exploration. The qualitative data suggest that the examiners in this study were able to assess equally well in both modes but that attaining this level of performance on screen exacted a greater cognitive workload on them. This finding mirrors those of other screen reading studies which suggest that reading on screen is cognitively more demanding than reading on paper (e.g. Wästlund *et al*., 2005). It also appears that the examiners were less able to spatially encode the

information accessed on screen compared with paper and that this contributed to them having a weaker mental representation of the text. Again, literature can be found which suggests this to be an unsurprising finding (e.g. Dillon, 1994; O'Hara and Sellen, 1997; Piolat *et al*., 1997). Most importantly, the qualitative analyses in this study help to explain the basis of this modal difference. Examiners' reading navigation styles and elements of their annotating behaviours differed substantially across modes and theory suggests that these differences are important because navigation and annotating can support readers in the process of building stronger mental representations.

Although this study suggests that examiners appeared to work harder on screen to achieve similar outcomes to paper marking, there are two key elements which might help to illuminate this relationship further. First, it is possible that the examiners attained similar levels of consistency across modes because they had enough spare cognitive capacity to accommodate the additional cognitive load exacted by the marking task in the screen environment. This suggests that in this study the examiners were still working below the threshold at which the cognitive effort was manageable enough to maintain currently acceptable levels of consistency. Secondly, the major factor which contributed to this heightened cognitive load in the screen marking environment related to frustration, with the novelty of the screen marking experience factoring heavily into this. Importantly, this factor had a transient quality, becoming clearly less important throughout the marking period as the examiners became more familiar with the experience.

A recommendation of this project is that future research should continue to explore how the characteristics of on-screen marking environments might affect examiner cognitive load and to explore whether there exists a point beyond which additional cognitive load might lead to unacceptable levels of marking consistency. Such a study might consider whether any mode-related marking effects exist when more examiners (with differing levels of expertise) mark a greater number of scripts which are lengthier, and include a wider diversity of characteristics.

References

Battiste, V. & Bortolussi, M. (1988). *Transport pilot workload: a comparison of two objective techniques*. Proceedings of the Human Factors Society 32nd Annual Meeting, 24–28 October, Anaheim, CA, 150–154.

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.

Dillon, A. (1994). *Designing Usable Electronic Text*. London: Taylor & Francis.

Emerson, L. & MacKay, B. R. (2006). Subjective cognitive workload, interactivity and feedback in a web-based writing program. *The Journal of University Teaching and Learning Practice*, **3**, 1, 1–14.

Fowles, D. (2008). *Does marking images of essays on screen retain marker confidence and reliability?* Paper presented at the International Association for Educational Assessment Annual Conference, 7–12 September, Cambridge, UK.

Hansen, W. J. & Haas, C. (1988). Reading and writing with computers: a framework for explaining differences in performance. *Comm. ACM*, **31**, 9, 1080–1089.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: North Holland Press, 239–250.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L. & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human Factors,* **34**, 429–439.

Hornbæk, K. & Frøkjær, E. (2001). *Reading electronic documents: the usability of linear, fisheye, and overview+detail interfaces*. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI Letters, **3**, 1, 293–300.

Hsieh, G., Wood, K. R. & Sellen, A. (2006). *Peripheral display of digital handwritten notes*. CHI 2006 Proceedings, Montreal, Quebec.

Jankowicz, D. (2004). *The Easy Guide To Repertory Grids*. Chichester: John Wiley & Sons.

Johnson, M. & Greatorex, J. (2008). Judging text presented on screen: implications for validity. *E-Learning*, **5**, 1, 40–50.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge MA: Harvard University Press.

Just, M. A. & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn and Bacon.

Kelly, G.A. (1955). *The Psychology of Personal Constructs*. New York: Norton.

Kurniawan, S. H. & Zaphiris, P. (2001). *Reading online or on paper: Which is faster?* Proceedings of HCI International 2001. Mahwah, NJ: Lawrence Erbaum Associates.

Marshall, C. C. (1997). *Annotation: from paper books to the digital library*. Proceedings of the Second ACM International Conference on Digital Libraries; Philadelphia, Pennsylvania.

Mayes, D. K., Sims, V. K. & Koonce, J. M. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics*, **28**, 367–378.

Mills, C. B. & Weldon, L. J. (1987). Reading text from computer screens. *ACM Comput. Surv*. **19**, 4, 329–357.

Moroney, W. F., Biers, D. W. & Eggemeier, F. T. (1995). Some measurement and methodological considerations in the application of subjective workload and measurement techniques. *International Journal of Aviation Psychology*, **5**, 87–106.

O'Hara, K. (1996). *Towards a Typology of Reading Goals. Rank Xerox Research Centre Affordances of Paper Project Technical Report EPC-1996–107*. Cambridge: Rank Xerox Research Centre.

O'Hara, K. & Sellen, A. (1997). *A comparison of reading paper and on-line documents*. In: S. Pemberton (Ed.), Proceedings of the ACM Conference on Human Factors in Computing Systems, Atlanta, Georgia. ACM Press: New York, 335–342.

Paek, P. (2005). *Recent Trends in Comparability Studies*. PEM Research Report 05–05.

Piolat, A., Olive, T. & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology*, **19**, 291–312.

Piolat, A., Roussey, J-Y. & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, **47**, 565–589.

Schilit, B. N., Golovchinsky, G. & Price, M. N. (1998). *Beyond paper: supporting active reading with free form digital ink annotations*. Proceedings of CHI 98, Los Angeles, CA.

Stanton, N. A., Salmon, P.M., Walker, G. H., Baber, C. & Jenkins, D. P. (2005). *Human Factors Methods: A Practical Guide for Engineering Design*. Aldershot: Ashgate Publishing.

Wästlund, E., Reinikka, H., Norlander, T. & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior*, **21**, 377–394.