CAMBRIDGE ASSESSMENT

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

Martin Johnson, Rita Nádas and Hannah Shiell

Paper presented at the British Educational Research Association annual conference,
Manchester University, September 2009

Background to the topic

Literature suggests that the transition from paper- to computer-based testing cannot be taken for granted and that comparability between modes of testing and assessment needs to be established empirically (Paek, 2005). Measures of reliability are an important means of demonstrating the validity of computer-based assessment. Another priority is to explore whether mode has an influence on the qualitative features of performances attended to by assessors. Literature suggests that on-screen assessment can inhibit reading comprehension, implying a need to explore how mode might influence assessors' judgments about longer textual performances (Dillon, 1994; Hansen and Haas, 1988; Kurniawan and Zaphiris, 2001; Mills and Weldon, 1987; O'Hara and Sellen, 1997; Piolat et al., 1997; Wästlund et al., 2005).

Research questions

The aim of the research was to investigate whether examiners could mark digital images of a set of English Literature essays as reliably on screen as they could in the traditional paper mode. It also considered whether examiners attended to the same essay features in both modes. The study investigated modal effects on examiners' cognitive load and their ability to spatially encode information whilst reading, since spatial encoding is theorised as an important aspect of reader comprehension building. The study also gathered quantitative and qualitative evidence of examiners' reading navigation behaviours and annotating strategies to augment the other findings.

Research methods

180 GCSE English Literature essay scripts were divided into two matched samples of 90 scripts. The scripts were then blind marked to establish a gold standard mark for each script. 12 examiners, chosen because of the high quality of their past marking, then marked one sample of scripts on paper and the other on screen.

In order to control the order of sample marking and marking mode, the examiners were allocated to four different marking groups with similar mean examiner performance ratings for each group. 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. The normal conditions of marking were replicated as much as possible.

To investigate essay marking reliability, examiners' marks were statistically compared across both modes and with an independent reference mark for each essay. To consider whether mode affected the script features (or constructs) being attended to by the examiners, Kelly's Repertory Grid technique (Kelly, 1955) was used to elicit constructs and ratings from two senior examiners. These were then used to build a profile of the scripts. The marking reliability analyses were then used to infer any potential relationship between construct recognition and mode.

UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

Cambridge Assessment is the brand name of the University of Cambridge Local
Examinations Syndicate, a department of the University of Cambridge.
Cambridge Assessment is a not-for-profit organisation.

Examiners’ cognitive load was measured during task completion using the National Aeronautics and Space Administration Task Load Index (Hart and Staveland, 1988). This enabled a comparison of each marker’s cognitive workload whilst marking in each mode.

In order to compare spatial encoding across modes, a sample of examiners were asked to recall the location of specific information from within specific essays. Scores were then calculated for examiners’ recall accuracy and these were compared across modes.

Examiners’ navigation flow and annotation practices were coded, observed and compared across a sample of scripts marked in both modes. These observations were then used to inform a series of semi-structured interviews with each examiner.

Analytical frame

The study used a mixed methods approach, seeking to complement the quantitative data analyses through the use of qualitative and interpretative methods. The use of ethnographic observation methods to gather contextualised data and inform follow-up semi-structured interviews was designed to enhance the cognitive psychological perspective supported by the quantitative analyses.

Research findings and/or contribution to knowledge

Using a variety of methods this study elicits some of the complex reading behaviours that pertain to the assessment of extended texts. Examiners were able to replicate levels of reliability on screen similar to those found on paper. There was also no evidence that the examiners were attending to different construct features within the essays. Interestingly, it appeared that working on screen required greater cognitive effort, with the qualitative observation data suggesting a variety of reasons for this and implying areas for further research.

References

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

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

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

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.

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

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

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

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

Wästlund, E., Reinikka, H., Norlander, T. and 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