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

Bramley, T. & 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.

Crisp, V. & Johnson, M. (*in press*). The use of annotations in examination marking: Opening a window into markers' minds. *British Educational Research Journal*.

Greatorex, J (2004). Moderated e-portfolio project evaluation. Evaluation and Validation Unit, UCLES. Available online at: www.ocr.org.uk/OCR/WebSite/Data/Publication/E-Assessment%20Materials/Moderated_82372.pdf (accessed 25 August 2005).

Laming, D. (2004). *Human judgment: The eye of the beholder*. London: Thomson.

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.

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

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

Preston, J. A. & Shackelford, R. (1999). Improving on-line assessment: An investigation of existing marking methodologies. *ACM SIGCSE Bulletin,* **31**, 3, 29–32.

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

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

Wilmut, J. (1984). *A pilot study of the effects of complete or partial removal of marks and comments from scripts before re-marking them*. AEB Research Report RAC 315.

Judging learners' work on screen: Issues of validity

Martin Johnson and Dr Jackie Greatorex Research Division

Introduction

Current developments in Cambridge Assessment and elsewhere include assessors[1] marking digital images of examination scripts on computer, rather than the original scripts on paper, and judges marking and moderating digitally produced coursework on computer, rather than on paper. One question such innovations raise is whether marks from judgements made about the same work presented on computer and on paper are comparable.

Generally the literature concerning the on-screen marking of tests and examinations suggests that on-paper and on-screen scores are indeed comparable (e.g. Bennett, 2003; Greatorex, 2004), although Fowles and Adams (2005) report that differences have been found in studies by Whetton and Newton (2002), Sturman and Kispal (2003) and Royal-Dawson (2003).

Our concern in this discussion article is that even when double marking studies find high levels of agreement between marks for the same work judged in different modes, issues of validity might be masked. We are thinking of validity in terms of the cognitive processes of the assessor when reaching a judgement, judgement per se and how well these reflect the judgements that were intended when the assessment was devised.

Do assessors all use the same criteria/guidelines in the same way?

If assessors do not use the same criteria/guidelines in the same way then validity is threatened. Laming (2004) argues that all judgements are comparisons of one thing with another and this process is influenced by the prior experience of the judge and the context in which the comparisons are being made. In the case of examination marking he explains that sometimes the mark scheme might leave room for examiner uncertainty, especially when marking essays, when mark schemes might be interpreted in different ways.

Other research evidence has suggested that assessors do not mechanistically match learners' achievement to assessment criteria (Wolf, 1995) and that a range of extraneous factors can influence assessors' decisions (Ragat and Hevey, 1995). These might include a respondent's handwriting (Green *et al*., 2003; Milanovic *et al*., 1996, Sanderson, 2001), the context in which an assessment is carried out (Wolf, 1995), or the assessor's own idiosyncratic internalised standards (Eraut *et al*., 1996).

Although the above research review is partial, it is intended to illustrate that human judgement in conventional assessment practices is potentially influenced by a number of extraneous variables, partly explaining why assessment criteria might be interpreted by different judges in different ways. We would like to explore how the mode of marking/moderating might also lead to the influence of extraneous variables.

[1] We use the terms assessor/judge as general terms to refer to examiners, moderators, verifiers or others who assess candidates' work.

If assessors are presented with the same work in different modes do they make qualitatively different judgements?

We are particularly interested in the area of assessment that is likely to place a high cognitive demand on the assessor, for example, the assessment of extended prose. It is possible that presenting such work on screen might affect how assessors read and perceive the text, and therefore, possibly, their assessment judgements.

The concept of affordance recognises that the environment influences subjects' behaviour, with some environments facilitating or inhibiting certain types of activity. Gibson (1979) suggested that part of the success of human evolutionary development has been a consequence of their ability to identify and exploit the affordances of different environments. He claimed that humans perceive affordance properties of the environment in a direct and immediate way and they subsequently perceive possibilities for action. In this sense, the modes of paper and computer exist as environments within which activity is carried out, and each has its own affordances.

An interesting implication of Gibson's theory concerns the evolution of behaviour as a response to environmental change. Using the Piagetian notion of learning as a function of assimilation and accommodation, Reinking *et al*. (2000) suggest that a lag may exist between technological developments and the accommodation of those changes as reflected in the behaviour, and specifically the reading and writing activities, of people interacting with that new technology. This implies that patterns of behaviour prevalent when working in a paper-based environment may continue even though the terrain of that environment has been changed by new computer technology.

Sellen and Harper (2002) compare the affordances of paper and of digital reading technologies.

Paper	*Digital reading technologies*
● Able to support flexible navigation	● Able to store and access large amounts of information
● Able to support cross document use	● Able to display multimedia documents
● Able to support annotation while reading	● Enable fast full-text searching
● Able to support the interweaving of reading and writing	● Allow quick links to related materials
	● Allow content to be dynamically updated or modified

Sellen and Harper (2005) suggest these affordances are linked to the purposes of reading and text length. Where the focus of reading only requires a superficial skim of a longer text, or the deeper processing of shorter text (i.e. text that can be confined to one screen at a reasonable resolution), mode-related effects on reading might be minimal.

A number of studies have investigated mode-related marking effects. Sturman and Kispal (2003) studied differences in Key Stage 2 spelling, reading and writing marks given when the same work was marked by examiners using a computer package to view the work and record the score and when it was marked on paper. Although there was an absence of any consistent trend, the differences between mean marks might suggest e-marking and conventional marking judgements were qualitatively different. Findings from studies by Price and Petre (1997) and Greatorex (2004) have identified a number of mode-related aspects that might qualitatively influence assessment judgements.

Page and document management

Sellen and Harper (2002) suggest that digital technologies have significant constraints when dealing with document layout, largely because any document needs to be presented within a particular screen size, which limits how many pages or documents can be made visually available at once. They contrast this with page and document management techniques on paper, where navigation is two-handed and allows the simultaneous overlap and interleafing of pages and documents. These techniques allow pages to be manoeuvred so that visual connections can be made, affording the reader a broader field of view than is possible on screen. Both Greatorex (2004) and Price and Petre (1997) find evidence to suggest that mode might affect judgement processes. In the Greatorex study moderators suggested that reading and navigating through e-portfolios was hindered because scrolling backwards and forwards was difficult. Viewing different candidates' work was another source of difficulty. In some situations moderators intended to almost simultaneously search for evidence in different pieces of work by the same candidate, but could not do this on screen as there seemed much more of a time lag between reading one piece and another. Where candidates had used more than one file to present their work, dipping in and out of multiple files to identify where the teacher had given credit was also burdensome and time-consuming. Price and Petre also reported that markers in their study found opening and switching between documents to be onerous.

Technological mediation of the text

Technology might draw an assessor's attention to information to which they would not otherwise have attended. Greatorex reports that some moderators noted that many of the e-portfolio files were in Microsoft Word™ which incorporated an automatic spell and grammar check, underlining spelling and possible grammar errors in candidates' work. These moderators suggested that this software made it easier to see Quality of Written Communication (QWC) errors on screen compared with the paper version of the same portfolios. The process involved in making the judgement about the e-portfolio and the paper version of the same portfolio were qualitatively different. This is an important issue since it reiterates the crucial relationship between assessment tools, assessment purpose and underlying validity.

Sense of text

Greatorex (2004) reported moderators trying to find their way around candidates' work to gain a sense of meaning from the text. E-moderation required moderators to scroll to see all the evidence for an assessment criterion, affording them the opportunity to see the information in snapshots with a limited view of the whole text. In contrast, when using paper portfolios they could glance from one area to another whilst maintaining a view of the broader text.

A wealth of research exists to suggest that spatial encoding takes place during the reading process and that this is an integral part of building a mental representation of the location of textual information (Piolat *et al*., 1997; Fischer, 1999; Kennedy, 1992). Piolat *et al*. cite a number of studies which suggest that paper supports the process where readers assign geographical locators to words and ideas during reading. It is inferred that this is not equally afforded by both paper and computer-based texts, in turn having implications for a reader's cognitive load. Pommerich (2004) has found evidence to suggest that readers' positional memory is better on paper because it appears that they can operationalise positional memory more easily. Reasons for this appear to relate to the fact that only having a limited view of a text on screen disturbs the reader's representation of its spatial layout and disrupts their construction of a reliable mental representation of the text. Research by Dyson and Haselgrove (2000) has also found evidence of subjects' extremely poor on-screen reading performance on structural questions that require the locational recall of at least two pieces of information. One explanation for this relates to differences in navigability around texts. Pommerich suggests that the scrolling techniques afforded by computer-based texts only allow relative spatial orientation since the field of vision is less than when navigating around paper-based texts. Furthermore, the relative slowness and imprecision of scrolling compared with manual navigation may lead to increased cognitive demands on the reader whilst they find their way around the text since this 'does not provide…enough "tangible" data about the location of information that is not currently on the screen. Each time the scroll arrows are used, or even the scroll bar, the spatial layout is disrupted', and moreover, 'scrolling through the text to find a particular piece of information can be slow' (Piolat *et al*., p.568). To conclude, Piolat *et al*. argue that the combination of slow and imprecise navigation around a text, disrupted spatial layout and the limited view of the text on screen make it difficult for a 'sense of text' to be constructed when accessed on computer.

Such findings imply that it is more cognitively demanding to gain a sense of text when e-moderating or e-marking than when assessing on paper.

Reading strategies

Greatorex (2004) reported that mode somewhat influenced how moderators read candidates' work. Teachers and moderators search through the portfolio to look for particular information to satisfy the assessment criteria. Some moderators reported that in reading paper portfolios they had spotted evidence that appeared to have been missed by teachers and moderators who previously assessed the electronic version of the same portfolios on screen.

O'Hara (1996) described a series of reading strategies. He found that (1) reading strategy choices are related to the purpose of the reading, for example, proof reading requires different strategies to reading for information, and (2) mode has a greater influence on reading strategies than the purpose of the reading. Askwall (1985) described a number of search strategies used by readers and showed that search strategy choices were influenced by mode. The results of Greatorex (2004) (mentioned above) are in keeping with this research literature. Therefore, the information gleaned by assessors about candidates' work and the sense they make of it might be affected by mode.

Annotation

Price and Petre (1997) found that assessors in their study had different marking styles on paper and on screen, but that some markers were more affected by mode than others. They also found that annotations used in paper and e-marking were different despite being available in both modes. Greatorex found that when teachers marked on screen they reported difficulties annotating directly onto candidates' work and said there would have been more annotations if the portfolios had been on paper.

Annotating can help text comprehension and affects how meaning is processed (O'Hara, 1996). In some circumstances it supports markers' decision-making processes (Crisp and Johnson, *in press*). Making paper-based annotations is relatively effortless and is part of the meaning construction process during reading but computer-based annotation might be impeded by a lack of usable annotation tools (O'Hara and Sellen, 1997). A number of recent developments have been designed to overcome these weaknesses, including stylus entry directly onto a tablet or touch screen, and digital ink technologies.

We deduce from the research literature that annotation plays a crucial role in text comprehension and that in some situations this might be important when making assessment judgements.

Conclusions

Although the above research review is partial, it is intended to illustrate that judgements are potentially influenced by a number of extraneous variables. There is some evidence that mode does not generally affect scores across different modes but that judgements are sometimes affected qualitatively, in which case validity can be enhanced or compromised.

If it is the case that mode affects assessment judgements, it must be considered in the wider context. First, it is just one of many influences, and secondly, the benefits of technology in education and assessment are well rehearsed in a large body of research literature (e.g. Salmon, 2004; Heppell, 2003; Sellen and Harper, 2002).

References

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

Bennett, R. E. (2003, November). *On-line assessment and the comparability of score meaning* (ETS Publication, RM-03-05). Retrieved August 31, 2005 from http://ftp.ets.org/pub/res/researcher/RM-03-05-Bennett.pdf

Crisp, V. and Johnson, M. (*in press*). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*.

Dyson, M. C., and Haselgrove, M. (2000). The effects of reading speed and reading patterns on the understanding of text read from screen. *Journal of Research in Reading*, **23**, 2, 210–23.

Eraut, M., Steadman, S., Trill, J. and Parkes, J. (1996, November). *The assessment of NVQs*. (Research Report No. 4) Brighton: University of Sussex Institute of Education.

Fischer, M. H. (1999). Memory for word locations in reading. *Memory*, **7**, 1, 79–118.

Fowles, D. and Adams, C. (2005). *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the IAEA Conference, Abuja, Nigeria. Retrieved February 5, 2006 from www.iaea.info/abstract_files/paper_051218101528.doc

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Greatorex, J. (2004, December). *Moderated e-portfolio project evaluation*. (Evaluation and Validation Unit, University of Cambridge Local Examinations Syndicate). Retrieved August 31, 2005 from http://www.ocr.org.uk/OCR/WebSite/Data/Publication/E-Assessment%20Materials/Moderated_82372.pdf

Green, S., Johnson, M., O'Donovan, N. and Sutton, P. (2003). *Changes in key stage two writing from 1995 to 2002*. Paper presented at the United Kingdom Reading Association Conference, University of Cambridge, UK, July 2003.

Heppell, S. (2003). Assessment and new technology: New straitjackets or new opportunities. In C. Richardson, (Ed.), *Whither Assessment?* London: Qualifications and Curriculum Authority. Retrieved August 31 2005 from http://www.qca.org.uk/downloads/combined_whither_assessment.pdf

Kennedy, A. (1992). The spatial coding hypothesis. In K. Rayner (Ed.), *Eye movements and visual cognition*. New York: Springer-Verlag.

Laming, D. (2004). *Human judgment: The eye of the beholder*. London: Thomson.

Milanovic, M., Saville, N. and Shuhong, S. (1996). A study of the decision-making behaviours of composition markers. In M. Milanovic and N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge*. Studies in Language Testing 3. Cambridge, UK: Cambridge University Press/UCLES.

O'Hara, K. (1996). *Towards a typology of reading goals*. (Rank Xerox Research Centre Affordances of Paper Project Technical Report EPC-1996-107). Cambridge, UK: Rank Xerox Research Centre.

O'Hara, K. and 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.

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**, 565–89.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based texts. *The Journal of Technology, Learning and Assessment*, **2**, 6. Retrieved August 31, 2005 from http://www.bc.edu/research/intasc/jtla/journal/pdf/v2n6_jtla.pdf

Price, B. and Petre, M. (1997). *Teaching programming through paperless assignments: An empirical evaluation of instructor feedback*. Centre for Informatics Education Research, Computing Department, Open University, UK. Retrieved April 20, 2005 from http://Mcs.open.ac.uk/computing papers/mzx/teaching.doc

Ragat, P. and Hevey, D. (1995). *Sufficiency of evidence: The development of guidance for Assessors and Verifiers*. (Research and Development Report No. 32), Sheffield, UK: DFEE.

Reinking, D., Labbo, L. D. & McKenna, M. C. (2000). From assimilation to accommodation: A developmental framework for integrating digital technologies into literacy research and instruction. *Journal of Research in Reading*, **23**, 2, 110–22.

Royal-Dawson, L. (2003). *Electronic marking with ETS software*. AQA Research Committee. Paper RC/219. In D. Fowles and C. Adams (2005), *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the IAEA Conference Abuja, Nigeria. Retrieved February 5, 2006 from www.iaea.info/abstract_files/paper_051218101528.doc

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

Sanderson, P. J. (2001). Language and differentiation in examining at A level. PhD thesis, University of Leeds.

Sellen, A. and Harper, R. (2005). Personal correspondence.

Sellen, A. and Harper, R. (2002). *The myth of the paperless office*. Cambridge, MA: MIT Press.

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 Annual Conference, Manchester, UK, October 2003.

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

Wolf, A. (1995). *Competence-based assessment*. Buckingham, UK: Open University Press.

NEW TECHNOLOGIES

The Cambridge Assessment/Oxford University automatic marking system: Does it work?

Nicholas Raikes Research Division

In the first issue of *Research Matters*, Sukkarieh *et al*. (2005) introduced our work investigating the automatic marking of short, free text answers to examination questions. In this article I give details and results of an evaluation of the final prototype automatic marking system that was developed.

Introduction

Background

Cambridge Assessment funded a three year research project that investigated the application of computational linguistics techniques to the automatic marking of short, free text answers to examination questions. The research, conducted at Oxford University by Professor Stephen G. Pulman and Dr. Jana Z. Sukkarieh, focussed on GCSE Biology as a suitable context since the Biology question papers contained large numbers of questions requiring short, factual, written answers.

The researchers took two broad approaches to automatic marking. The first approach involved writing by hand what were, loosely speaking, machine marking schemes for the items to be automatically marked. This approach is referred to as the 'Information Extraction' approach. The second approach – dubbed the 'Machine Learning' approach – involved trying various machine learning techniques to, again loosely speaking, learn the marking scheme from a sample of human marked answers. A hybrid approach using semi-automatic methods to produce