References

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, **7**, 31–51.

Ericsson, K. & Simon, H. (1993). *Protocol Analysis: Verbal reports as data*. London: MIT Press.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgments. *Quarterly Journal of Experimental Psychology*, **42A**, 239–54.

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

Leighton J. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, Winter, 6–15.

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

Stanovich, K. & West, R. (2002). Individual differences in reasoning. In T. Gilovich, D. Griffin & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Suto, W.M.I. & Greatorex, J. (*in press*). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*.

Van Someren, M., Barnard, Y. & Sandberg, J. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London: Academic Press.

Vaughan, C. (1992). Holistic assessment: what goes on in the rater's mind? In L. Hamp-Lyons (Ed.). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.

Webster, F., Pepper, D. & Jenkins, A. (2000). Assessing the undergraduate dissertation. *Assessment and Evaluation in Higher Education*, **25**, 71–80.

Yorke, M., Bridges, P. & Woolf, H. (2000). Mark distributions and marking practices in UK higher education. *Active Learning in Higher Education*, **1**, 7–27.

Examiners' annotations: Practice and purpose

Victoria Crisp and Martin Johnson Research Division

Introduction

> *'When you come to any passages that seem to you useful, make a firm mark against them, which may serve as lime in your memory, less otherwise they might fly away.'*
> Advice from St Augustine in Petrarch: *Secretum Meum* 1358

The processes of reading and writing are recognised to be inextricably intertwined. Writing helps to support cognitive demands made upon the reader whilst processing a text (e.g. O'Hara, 1996; Benson, 2001). Anderson and Armbruster (1982) suggest that annotating activities are concurrent with the actual reading processes, influence the way that reading occurs, and the way that meaning is processed. Examiners annotate scripts whilst marking (e.g. underlining, circling, using abbreviations or making comments) and this may reflect the cognitive support for comprehension building that annotations can provide.

Within the accountability agenda that pervades education there is an emphasis on clear communication channels between examiners of different seniority to facilitate effective monitoring. Annotations might have an important communicative role in this quality control process by offering others up and down the chain an insight into the rationale behind the annotating examiners' decisions. Previous re-marking investigations have suggested that annotations do have a communicative function, potentially influencing how subsequent viewers perceive the quality of a script (Murphy, 1979; Wilmut, 1984; Newton, 1996). Laming (2004) suggests that this is because there are places where the mark scheme leaves the examiner uncertain, and that judgements in such cases are influenced by extraneous information, for example, the previous annotations of other judges.

In addition to evidence that annotations act as a communicative device, there is also evidence that annotating might have a positive influence on markers' perceptions and affect their feelings of efficacy. Most markers felt that annotating improved their marking, helping them to apply performance criteria and reducing the subjectivity of judgements (Bramley and Pollitt, 1996). In pilot work on online assessment teachers, examiners and moderators have expressed dissatisfaction where facilities for annotation were limiting (Greatorex, 2004; Raikes *et al.*, 2004). Markers report that using annotations provides an efficient means to confirm or reconsider standards both within and across candidates as well as acting as a reassurance during the judgemental process (Shaw, 2005).

Rationale

The literature available provides some information about the purposes and effects of annotations. However, there is a relative sparsity of published research about annotation in examination marking in terms of the following:

- consistency of use of codes
- examiners' reasons for using annotations
- the role that annotations might be playing in decision making processes
- the effects, or perceived effects, of using annotations whilst conducting first marking.

This research investigates some of these issues and develops a more comprehensive picture of annotation practices.

Annotations in different subjects

Before the main study, we analysed some marked examination scripts to build a basic profile of annotation use across a variety of subject disciplines. Three scripts from each of sixteen subjects were analysed. The types of annotation used and the frequency of their use varied substantially between subjects. However, six types of annotation were used fairly frequently across a number of subjects:

- making comments
- underlining
- sidelining
- circling
- question marks
- carets.

There were no significant relationships between frequency of annotation and subject families. However, analysis did show a significant relationship between the length of response and mean number of annotations ($p = 0.007$) with comments, in particular, being more common where questions require longer answers.

Main study method

The main study investigated one Mathematics GCSE and one Business Studies GCSE paper. Six examiners (including the Principal Examiner) who had previously been standardised to mark the paper were recruited for each subject.

Examiners initially marked ten scripts which were then reviewed by their Team Leader. Examiners then marked a further 46 (Business Studies) or 40 (Mathematics) scripts. A representative sample of scripts was used. The scripts used were photocopies of the original scripts with all marks and annotations removed to simulate a first marking situation. Each examiner within a subject marked the same candidates' scripts. The reliability of the experimental marking was investigated and analyses suggested that marker behaviour in this study was generally representative of marker behaviour during live marking (see Crisp and Johnson, *in press*, for details).

The examiners later attended individual meetings with researchers. The session began with each examiner marking a small number of new scripts to re-familiarise themselves with the examination paper and mark scheme. A researcher then observed each examiner as they continued marking a few further scripts. This recorded whether individual annotations were made before or after the recording of marks (for details and findings of this stage see Crisp and Johnson, *in press*). Later in the session each examiner was interviewed about their use of annotations.

Analysis of annotation practices

Annotation use in the two subjects differed in a number of ways (see tables below). Business Studies markers used roughly twice as many annotations per question part as Mathematics markers. Furthermore, the most common types of annotations used were different for the two subjects. However, in both subjects markers varied in the frequency and types of annotations they used compared with others in their team, with no obvious relationship found between marker reliability and annotation use between markers. For Business Studies there were significant differences between the frequencies of use of most annotations across examiners, perhaps reflecting different preferences or habits. For Mathematics, despite sometimes significant variations between markers in frequency of annotation use, there was an underlying similarity in the types of annotations used.

Mean number of annotations per script (Business Studies)

Examiner	*BOD*	*Underline*	*Comment*	*Sideline*	*Level*	*Question Mark*	*All annotations*
1	0.83	0.00	0.00	0.00	0.00	0.00	0.91
2	0.87	2.96	2.78	2.61	3.43	3.00	17.52
3	1.70	1.61	0.48	1.30	8.52	0.13	14.96
4	0.43	3.96	5.22	0.09	5.96	0.91	17.87
5	0.04	0.65	0.39	1.70	11.83	0.22	15.87
6	0.35	6.39	0.52	0.26	9.52	0.17	17.48
All examiners	0.70	2.59	1.57	0.99	6.54	0.74	14.10

Mean number of annotations per script (Mathematics)

Examiner	*Circle*	*Underline*	*Comment*	*Caret*	*M*	*Sideline*	*All annotations*
1	1.35	0.95	0.55	1.60	1.15	0.00	6.70
2	1.00	1.05	0.70	2.45	2.50	0.50	10.00
3	2.15	1.35	0.95	1.90	2.40	1.53	12.26
4	1.05	1.10	0.10	3.30	1.75	0.45	8.95
5	1.85	0.55	0.95	2.00	1.75	0.65	8.85
6	1.40	4.05	0.95	2.40	2.25	0.15	13.00
All examiners	1.47	1.51	0.70	2.28	1.97	0.54	9.94

Note: Some of the less frequently used annotation types have been omitted from these tables

In both subjects some markers sometimes employed their own stylistic annotations in preference to commonly recognised standard annotations used by others. For example, instead of using circles to indicate errors one marker chose to place errors in parentheses.

Analysis of interviews

Interviews were carried out to probe examiners' ideas about the purposes of annotating, where examiners gained their annotation knowledge, and to elicit their perceptions of annotating.

Using annotations to justify decisions to others appeared to be the most salient purpose for all Mathematics markers. The Business Studies examiners also reported justifying or explaining decisions to be one of the purposes of annotation and placed additional emphasis on communicating to other examiners the reasoning behind awarding marks. One examiner said, *'It's a way of telling somebody else why and when I reach a decision. Er, whether they agree with it or not, at least they can see where I'm coming from.'*

Examiners also reported that annotating supported their judgements (particularly where marginal decisions had to be made) and helped them reach the appropriate mark. For example, one examiner said, '*I feel it's there to help the examiner reach the correct mark for the question; so it's an aid*'. Another said, '*So my annotation really is my way of thinking aloud but without talking aloud. Er, it's my way of communicating to myself.'*

Most markers felt that annotating helped to structure their thinking whilst marking, particularly for questions where candidates needed to

show their working or where multiple marks were available in Mathematics or when using levels of response marking in Business Studies.

All examiners reported that the annotations they used came from guidance in the mark scheme or (in Business Studies) from an information sheet provided separately. These were often used to support brief discussion at standardisation meetings. Annotation conventions were often reported to be well established and commonly understood among markers, with relatively little room for markers to use non-standard annotations. This was especially the case amongst Mathematics examiners. Half of the Mathematics markers suggested that the community's annotation conventions were based *'pretty well in folklore'*, *'word of mouth'* or *'custom'*. Other markers conveyed a profile of a centralised marking community with clear guidance about how to apply conventions. Two Business Studies examiners reported that their annotations were based on what the Team Leader and other examiners used. Those Business Studies examiners who gave a view thought that the types of annotations used were fairly consistent within the examining team although patterns and frequency of use might vary.

However, in both subjects there was space for marker idiosyncrasies provided that it did not compromise common understanding. Some examiners explained that they felt it necessary to sometimes use full words and sentences rather than standard annotations in order to communicate their reasoning clearly.

All but one marker felt that annotating was a positive aspect of marking. Again, the dual functions of accountability (*'It's useful for the examiner, the chief examiner, to know how you've allocated the marks'*) and supporting marker judgements (*It gives more clarity in your own mind when you're awarding marks'*) were the most clearly stated reasons used to support this. The examiner who viewed annotation negatively expressed two reasons for this opinion: first that he/she was not *'fully aware of how to use them to their best effect'* and secondly that he/she did not see the need to use them saying that the *'mark scheme is what I'm marking to'*.

Discussion

The findings portray a clear sense that markers in both subjects believed that annotating performed two distinct functions. The first appeared to be justificatory, communicating the reasons for their marking decisions to others. This mirrors the statutory requirements for awarding bodies to establish transparent, accountable procedures which ensure quality, consistency, accuracy and fairness. The second purpose was to support their thinking and marking decisions. In addition to helping markers with administrative aspects of marking (for example, keeping a running tally of marks) there are claims that annotations also support higher order reading comprehension processes.

There are also suggestions that annotations can help to provide a 'visual map' of the quality of answers (Bramley and Pollitt, 1996). This is perhaps especially useful for the purpose of making comparisons – especially between longer texts that possibly exact a great deal of cognitive demand on the marker. Laming (2004) suggests that making comparisons is a key element in the process of making judgements. So it is not surprising that the notion of 'annotating to support thinking' appeared to be more salient for Business Studies markers, who were more likely to deal with longer answers, than for Mathematics markers.

From the study it appears that different subjects have different annotation profiles. There seems to be a 'pool' of annotations pertinent to each subject, and structures exist within subject marking communities to transmit information about appropriate annotation use, for example, through mark schemes and standardisation meetings.

This said, individual markers dipped into their subject 'annotation pool' in different ways. It was common for markers to have their own particular annotation profile, using different subsets of available annotations and using certain annotations with varying frequency. It was uncommon for examiners to refer to annotation lists whilst marking, suggesting that their annotations were internalised. Preston and Shackelford (1999) report similar findings in the context of assessing computer science. They found that raters did not refer to the list of feedback codes used to classify errors but that 'raters remember and use a small subset of all available (and appropriate) feedback codes while marking' (p. 31). This variation between examiners could seem negative but it is clear that the use of annotation is currently an efficient communicative practice that is an automatic part of the marking process for those embedded in the examining culture, and rarely something that is 'tagged on'. Examiners may vary in their usage but whatever they are doing it appears to support their work and they are positive about the role that it plays.

Despite room for marker idiosyncrasy the key underpinning feature of annotation use appeared to be that it needed to be commonly understood by other members of the community. This reflects the role of annotation as a communicative tool, reflecting notions embedded in Situated Learning Theory. This theory suggests that effective working communities are based around sets of common norms and practices. Effective communication between community members is essential to the efficient working of the group. Part of this communication might involve the evolution and use of specialised tools that facilitate the transmission of knowledge between community members. To some extent it appears that marker annotation practices conform to this model, behaving as communicative tools and carrying a great deal of meaning to those within the community. Study findings suggest that markers believe annotating to be a positive aspect of marking which concurs with other findings (Bramley & Pollitt, 1996).

This research has gathered evidence on current annotation practices in examination marking in two subjects. Whilst the data available do not suggest that patterns of annotation use dramatically affect marker reliability, the practice of annotation sometimes supports marker judgement, is generally viewed positively and appears to give examiners confidence in their professional judgements.

Acknowledgements

We would like to thank the examiners involved for their participation and for sharing their views with us. We would also like to thank Jackie Greatorex and John F. Bell for the statistical analysis relating to the reliability of the experimental marking.

Further reading

This report is a summary of an article, 'The use of annotations in examination marking: opening a window into markers' minds', which is *in press* at the *British Educational Research Journal*.

References

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

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.

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

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.