Acknowledgements

Jane Fidler has worked on some of the data management and presentation of images, graphs, figures and tables in this article. Rita Nadas has also worked on some of the data management in this study.

References

Greatorex, J. (2006). *Do examiners' approaches to marking change between when they first begin marking and when they have marked many scripts?* A paper presented at the British Educational Research Association Annual Conference, September 2006, University of Warwick.

Greatorex, J. & Suto, W. M. I. (2006). *An empirical exploration of human judgement in the marking of school examinations*. A paper presented at the International Association of Educational Assessment conference, May 2006, Singapore.

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.

Pinot de Moira, A., Massey, C., Baird, J. & Morrissy, M. (2002). Marking consistency over time. *Research in Education*, **67**, 79–87.

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 submission*). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations.

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*.

Suto, W. M. I. & Greatorex, J. (2006). A cognitive psychological exploration of the GCSE marking process. *Research Matters: A Cambridge Assessment Publication*, **2**, 7–10.

Researching the judgement processes involved in A-level marking

Victoria Crisp Research Division

Introduction

The marking of examination scripts by examiners is the fundamental basis of the assessment process in many assessment systems. Despite this, there has been relatively little work to investigate the process of marking at a cognitive and socially-framed level. Vaughan (1991) and others have commented on the importance of investigating the process and decision-making behaviour through which examiners make their evaluations. According to Milanovic, Saville and Shuhong (1996), the lack of understanding about the decision-making process makes it hard to train examiners to make valid and reliable judgements. A decade later their view is still accurate. Improved understanding of the judgement processes underlying current assessment systems would also leave us better prepared to anticipate the likely effects of various innovations in examining systems such as moves to on-screen marking.

The research summarised here started by reviewing the relevant literature in the areas of cognitive judgement, theories of reading comprehension, social theories of communities and research specifically investigating the decision-making and judgements involved in marking. Notable amongst the latter are the works of Vaughan (1991), Pula and Huot (1993) and Huot (1993) in the context of assessing writing, Milanovic, Saville and Shuhong (1996), Cumming (1990) and Lumley (2002) in the context of English as a second language, Sanderson (2001) with regard to marking A-level sociology and law essays and Greatorex and Suto (2006) in the context of short answer questions in maths and business studies GCSE papers. Few studies have researched the marking of disciplines other than English writing and none have considered the processes involved in marking short answer questions and essays within the same domain. This research was designed and undertaken to address this gap in our understanding of examiners' judgements and attempted to draw on a wider range of relevant theoretical areas than have been used in most previous studies.

Method

An AS level and an A2 level geography exam paper were selected. The AS level exam required students to provide short to medium length responses whilst the A2 unit involved writing two essays from a choice. Six experienced examiners who usually mark at least one of the two papers participated in the research. Five of the examiners were usually only involved in the marking of one of the papers but most had experience of teaching both units and would be eligible to mark the other.

Examiners marked fifty scripts from each exam at home with the marking of the first ten scripts for each reviewed by the relevant Principal Examiner. This reflected normal marking procedures as far as possible but the marking was not subject to the same degree of standardisation as live marking. Examiners later came to meetings individually where they marked four or five scripts in silence and four to six scripts whilst thinking aloud for each exam, and were also interviewed.

The scripts marked were photocopies of live scripts with marks and annotations removed. Examiners marked the same students' scripts, except that in the silent marking and think aloud marking, for each

examiner one of the scripts in each batch was a clean copy of one of the scripts included in the main batch of home marking.

Results

Analysis of the marks awarded during the home marking suggested that marking was broadly in line with live marking but that examiners tended towards severity in comparison. One examiner's marking of the AS unit was more severe than the others' and out of line with live marking and the same was the case for a different examiner's marking of the A2 unit.

The analysis of mark changes between home marking and silent marking at the meeting, and between home marking and marking whilst thinking aloud for the small number of repeated scripts suggested that thinking aloud affected the marks awarded very little, if at all. Thinking aloud seemed to result in slightly more consistent marking for short and medium length responses and slightly less consistent marking with essays, but these differences were small and could have occurred by chance. This helps to confirm that verbal protocol analysis is a valid research method in the investigation of the judgements involved in exam marking.

Coding the verbal protocols

Transcripts of the verbal protocols were analysed to try to understand the processes involved in the marking. Drawing on the transcripts and the work of Sanderson (2001) and Milanovic *et al*. (1996) a detailed coding frame was developed to reflect the specific qualities of student work noticed by markers and marker behaviours and reactions. The codes were grouped into the categories of:

- 'reading and understanding' (codes relating to reading and making sense of responses);
- 'evaluates' (codes relating to evaluating a response or part of a response);
- 'language' (codes relating to the student's use of language);
- 'personal response' (affective and personal reactions to student work);
- 'social perception' (social reactions such as making assumptions about candidates, talking to or about candidates, comments about teaching);
- 'task realisation' (codes relating to whether a student has met the demands of the task such as length of response, addressing/understanding question);
- 'mark' (codes relating to assessment objectives and quantifying judgements).

These categories are described in a little more detail below with short quotes from the verbal protocols included to exemplify the behaviours/reactions being described where this is helpful.

Reading and understanding

Not unexpectedly, reading and interpretation behaviours were frequent in the verbal protocols, perhaps emphasising the sometimes over-looked importance of reading and making sense of a student's intended meaning as a prerequisite to accurate and valid marking. Codes in this category included, among others, obvious reading behaviours, summarising or paraphrasing all or part of a response and seeking or scanning for something in particular in the student's work (e.g. '*really we are looking for two regions well described and explained to illustrate that unevenness*').

Evaluating

Also frequently apparent in the verbal protocols (and not unexpected) were behaviours relating to evaluating the text being read. Clearly positive and negative evaluations (e.g. '*so that's a good evaluation point', 'no she hasn't got the correct definition of site, she is confusing it*') were particularly frequent whilst other behaviours such as weighing up the quality of part of a response and making comparisons with other responses were also apparent.

Language

For both exam papers, all examiners made some comments about the quality of the language used by students. Some examiners also commented on the orthography (handwriting, legibility and general presentation) of student work, particularly with the essay paper (e.g. *'bit of a difficulty to read this towards the end, he has gone into scribbly mode*'). Comments on language and orthography were often negative.

Personal response

This category was created to accommodate the affective (i.e. emotional) reactions of some examiners to student work that sometimes occurred and other personal comments or responses. Reactions in this category included positive or negative affect (e.g. *'I quite like that', 'London [groan] my heart drops'*), laughter and frustration or disappointment. All examiners showed one or more of these reactions at some point but behaviours in this category were generally fairly infrequent except in the case of one examiner.

Social perception

Examiners sometimes displayed reactions associated with social perceptions of the imagined candidates. Markers sometimes made assumptions about the likely characteristics of the candidate (e.g. *'clearly not a very bright candidate'*), predicted further performance of the candidate (e.g. '*this is not going to be a better paper is it?'*) and talked to or about the candidate, sometimes almost entering into a dialogue with the student via the script (e.g. *'so give us an example now of this'*). Comments about teaching were also grouped into the category. Social perception type behaviours were generally fairly infrequent and varied in frequency between examiners, perhaps reflecting the personalities of individual examiners.

Task realisation

The comments coded in this category were about features of the responses required of students in order to successfully achieve the task set by a question and included comments on the length of responses, on material missing from a student's response (e.g. '*that doesn't really say why and it doesn't use map evidence*'), on the relevance of points made and on whether the candidate has understood and addressed the intended question.

Mark

A number of different types of behaviours relating to quantifying evaluations and making a mark decision were observed. These included (among other behaviours) comments regarding the Assessment Objectives stated in the mark scheme (particularly for the A2 exam),

initial indications of marks, reference to the mark scheme or to marking meetings or to the Principal Examiner's guidance and reflections on the total mark scored or on their own leniency or severity.

The following table shows a transcript extract from an examiner's marking of a short answer response along with the codes that were applied to this extract.

Transcript extract	*Codes*
Now we have got Mexico, Mexico city from rural areas, ok,	Summarises/paraphrases positive evaluation
increasing at a rate, mentions job opportunities, well explained there,	Summarises/paraphrases positive evaluation
a high standard, cramped housing, talking about what it is like in Mexico city rather than the whole country, (.)	Summarises/paraphrases neutral evaluation
shanty towns, now it's gone on to talk,	Summarises/paraphrases
most of it is irrelevant there,	Negative evaluation and relevance
but, let's have a look and see, explanation in only one area, [using mark scheme] (.)	Reference to mark scheme
so it's level 2 and is fairly general	First indication of mark
so I think we will give that 5	Mark decision
because it hasn't really explained much more than, not a lot about what it is like where they've come from, so really only explaining one area, southern	Discussion/review of mark/re-assessment

Findings

Did the frequencies of coded behaviours and reactions vary between the marking of different types of questions (short and medium length questions versus essays)?

The frequencies of codes were compared between the exam papers in order to consider whether there were differences in the behaviours involved in marking short to medium length responses and marking essays. There was no significant difference in the average total number of codeable behaviours per script between the two exams but there were a number of differences in the frequencies of individual codes. Differences included greater frequencies of two codes relating to social perceptions (assumptions about characteristics of candidates, predicting further performance) with the essay paper than with the AS exam. In addition, there were more instances of comments about addressing the question and about orthography (handwriting, legibility, presentation) with the A2 exam and greater acknowledgement of missing material with the AS exam. There were also differences in the frequencies of 'mark' related codes with more frequent reference to assessment objectives in the A2 exam, and more frequent occurrence of other mark related codes such as 'first indication of mark', 'discussion/review of mark/re-assessment' with the AS unit due to the greater number of mark decisions that have to be made. Examiners more frequently reflected on the total mark when marking the essay paper than with the shorter answer paper.

These differences give us some insight into the areas in which there might be a greater risk of examiner bias for each type of exam paper. There is more potential for assumptions about candidates or predicting performance in advance of a full reading to cause bias with essays than with shorter questions. There may be more risk of poor handwriting causing bias with essays. In addition, examiners are more likely to focus on what is missing from shorter responses than with essays. This is not to say that there was clear evidence of examiner bias occurring in these areas or that these are significant areas of risk but that these may be areas of potential risk worth bearing in mind when planning examination specifications and in marker training.

Did the frequencies of different types of behaviours and reactions vary between different examiners?

Differences between examiners in the frequencies of occurrence of codes were found for 31 of the 42 codes. Despite the variations in the frequencies of occurrence of individual behaviours or reactions between examiners, it seems that in most instances these differences did not have a significant impact on the agreement of marks between markers and that different marking styles can be equally effective.

Detailed analysis of the behaviours evidenced in the verbal protocols of the two examiners (one with the AS exam and one with the A2 exam) who awarded total marks that were significantly different to those of the other examiners offered some tentative hypotheses about influences on reliability. For example, greater frequencies of first indications of marks and discussion of marks were associated with lower marker agreement for one examiner which might suggest that over-deliberating on marking decisions is not advantageous. Lower frequencies of obvious reading behaviours were associated with lower marker agreement for both examiners, as were lower frequencies of comparisons with other scripts/responses and lower frequencies of positive evaluations.

Did the frequencies of coded behaviours and reactions vary between questions and/or between scripts?

Differences in the frequencies of code occurrence between questions were found for around half of the codes and were often associated with one particular essay question on a popular topic. There were few differences between scripts in the frequencies of codes that were applied suggesting that marking behaviours for different students' scripts were similar and that the findings are likely to be generalisable to other students' scripts beyond the sample used in the research. It seems that the processes involved in marking are infrequently affected by features of the scripts.

Which codes frequently occurred together?

Considering the frequently co-occurring codes also provided some interesting findings. Evaluations were often associated with aspects of task realisation (e.g. missing material, addressing/understanding question) and with the assessment objectives. Additionally, evaluations (especially negative evaluations) were often associated with considerations of the marks to be awarded. Positive evaluations and negative evaluations often co-occurred reflecting instances where examiners considered the strengths and weaknesses of a response or part of a response (e.g. *'a vague comment about the relief of the area'*).

Towards a model of the marking process

Analysis of the sequences of the coded behaviours apparent in the verbalisations allowed a tentative model of the marking process to be constructed. The model conceptualises three main phases and less frequently occurring 'Prologue' and 'Epilogue' phases before reading begins and after mark decisions have been made. The model attempts to

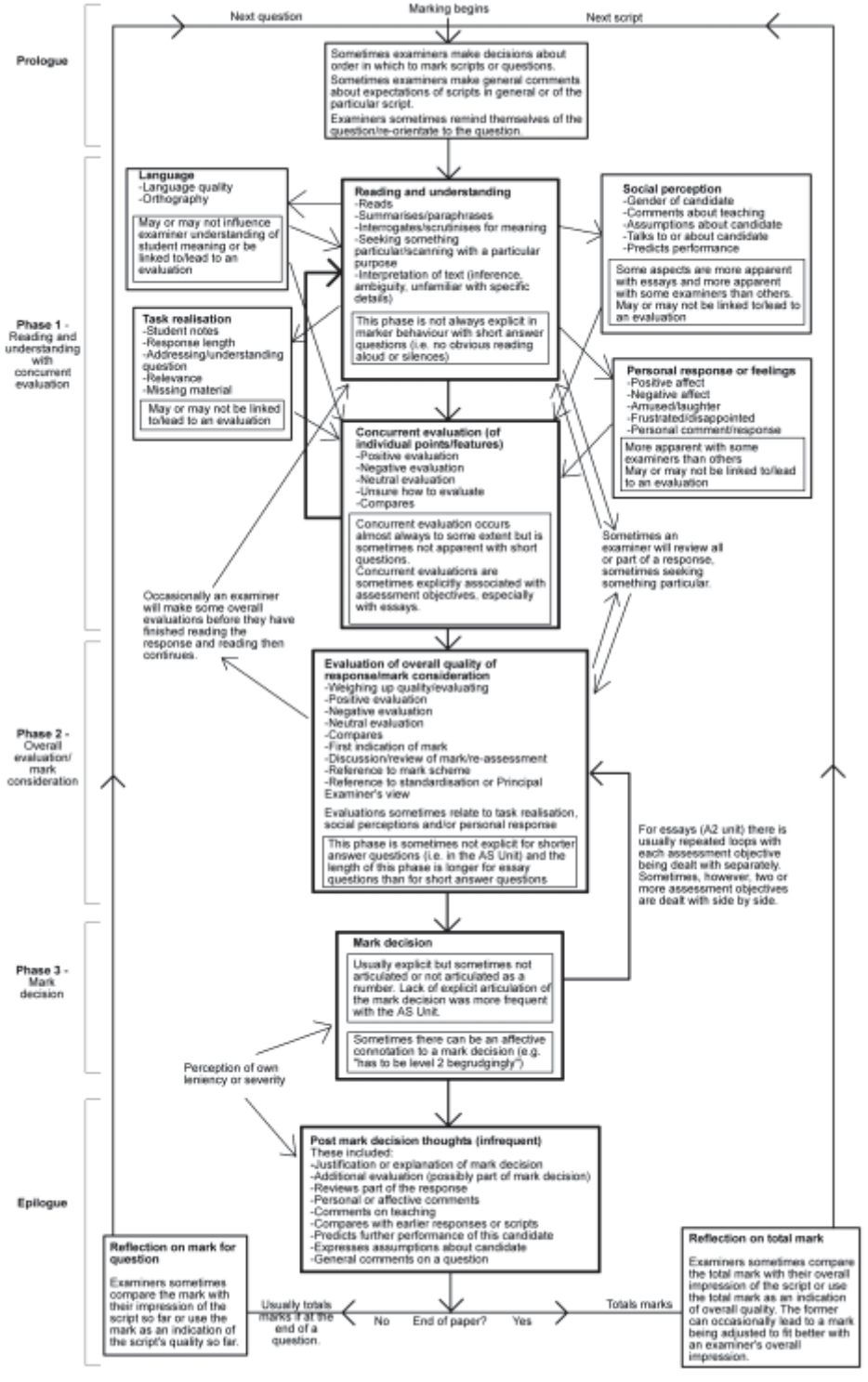
Prologue
Next question
Marking begins
Next script
Sometimes examiners make decisions about order in which to mark scripts or questions.
Sometimes examiners make general comments about expectations of scripts in general or of the particular script.
Examiners sometimes remind themselves of the question/re-orientate to the question.
Phase 1 - Reading and understanding with concurrent evaluation
Language
-Language quality
-Orthography
May or may not influence examiner understanding of student meaning or be linked to/lead to an evaluation
Reading and understanding
-Reads
-Summarises/paraphrases
-Interrogates/scrutinises for meaning
-Seeking something particular/scanning with a particular purpose
-Interpretation of text (inference, ambiguity, unfamiliar with specific details)
This phase is not always explicit in marker behaviour with short answer questions (i.e. no obvious reading aloud or silences)
Social perception
-Gender of candidate
-Comments about teaching
-Assumptions about candidate
-Talks to or about candidate
-Predicts performance
Some aspects are more apparent with essays and more apparent with some examiners than others. May or may not be linked to/lead to an evaluation
Task realisation
-Student notes
-Response length
-Addressing/understanding question
-Relevance
-Missing material
May or may not be linked to/lead to an evaluation
Personal response or feelings
-Positive affect
-Negative affect
-Amused/laughter
-Frustrated/disappointed
-Personal comment/response
More apparent with some examiners than others
May or may not be linked to/lead to an evaluation
Concurrent evaluation (of individual points/features)
-Positive evaluation
-Negative evaluation
-Neutral evaluation
-Unsure how to evaluate
-Compares
Concurrent evaluation occurs almost always to some extent but is sometimes not apparent with short questions.
Concurrent evaluations are sometimes explicitly associated with assessment objectives, especially with essays.
Sometimes an examiner will review all or part of a response, sometimes seeking something particular.
Occasionally an examiner will make some overall evaluations before they have finished reading the response and reading then continues.
Phase 2 - Overall evaluation/ mark consideration
Evaluation of overall quality of response/mark consideration
-Weighing up quality/evaluating
-Positive evaluation
-Negative evaluation
-Neutral evaluation
-Compares
-First indication of mark
-Discussion/review of mark/re-assessment
-Reference to mark scheme
-Reference to standardisation or Principal Examiner's view
Evaluations sometimes relate to task realisation, social perceptions and/or personal response
This phase is sometimes not explicit for shorter answer questions (i.e. in the AS Unit) and the length of this phase is longer for essay questions than for short answer questions.
For essays (A2 unit) there is usually repeated loops with each assessment objective being dealt with separately. Sometimes, however, two or more assessment objectives are dealt with side by side.
Phase 3 - Mark decision
Mark decision
Usually explicit but sometimes not articulated or not articulated as a number. Lack of explicit articulation of the mark decision was more frequent with the AS Unit.
Sometimes there can be an affective connotation to a mark decision (e.g. "has to be level 2 begrudgingly")
Perception of own leniency or severity
Epilogue
Post mark decision thoughts (infrequent)
These included:
-Justification or explanation of mark decision
-Additional evaluation (possibly part of mark decision)
-Reviews part of the response
-Personal or affective comments
-Comments on teaching
-Compares with earlier responses or scripts
-Predicts further performance of this candidate
-Expresses assumptions about candidate
-General comments on a question
Reflection on mark for question
Examiners sometimes compare the mark with their impression of the script so far or use the mark as an indication of the script's quality so far.
Usually totals marks if at the end of a question.
No
End of paper?
Yes
Totals marks
Reflection on total mark
Examiners sometimes compare the total mark with their overall impression of the script or use the total mark as an indication of overall quality. The former can occasionally lead to a mark being adjusted to fit better with an examiner's overall impression.

Figure 1 : (opposite)
A tentative model of the marking process in A level geography

bring together the various aspects of, and influences on, the marking process (text comprehension, cognitive processes, social perceptions, and personal responses) and is compatible with other research in the literature. Variations between marking short-answer questions and marking essays were apparent in certain phases of the model. The phases are outlined briefly below.

Prologue

When marking begins examiners sometimes make decisions about the order in which to mark scripts or questions and sometimes comment on their expectations of the scripts (e.g. *'surely we will have a good one soon'*) or re-orientate themselves to the question they were about to mark. The prologue occurs fairly infrequently.

Phase 1 – Reading and understanding with concurrent evaluation and comments on social perceptions of candidates, personal/affective responses and task realisation

This phase often involves loops of reading and/or paraphrasing parts of the response and then evaluating that part of the response. The process of making sense of the response and making concurrent evaluations tends to be less obvious with short answer questions. Concurrent evaluations are sometimes associated with assessment objectives, especially when marking essays. Reading a student's response can also trigger thoughts regarding the language used by the candidate, task realisation and social and personal responses, and these were sometimes directly associated with, or followed by, a concurrent evaluation.

Phase 2 – Overall evaluation/mark consideration

In phase 2 the examiner evaluates the response in a more overall way, possibly weighing up its quality, commenting on strengths and weaknesses. Explicit attempts are likely to be made at this stage to quantify the quality of the response with respect to the mark scheme. The examiner may have initial thoughts as to an appropriate mark and they may consider the advice in the mark scheme and given by the Principal Examiner during standardisation. The evaluations that are made at this stage may relate back to earlier thoughts regarding the task realisation, social perceptions and personal responses that impacted on concurrent evaluations. For the A2 exam, overall evaluations are usually made with regard to each assessment objective in turn and looping occurs between phases 2 and 3.

Phase 3 – Mark decision

This phase involves the examiner's decision about the mark. This was usually explicit in protocols but not always, particularly with short answer questions, perhaps because the mark decision occurs quickly and is consequently not articulated. Examiners sometimes reflected on the leniency or severity of their marking when deciding on a mark.

Epilogue

Fairly infrequently, additional consideration of the response occurs after the mark decision has been made. This can include, for example, justifying or explaining a mark decision, further evaluation, reviewing part of the response, personal or affective comments, comparisons with other scripts or responses, prediction of further performance by the candidate, and checking whether a total mark matched their overall impression of the script.

The tentative model is illustrated as a flow chart in Figure 1. The model requires further thought and development as well as validation in other subjects and assessments. The interview data were consistent with the coding frame and the proposed model of the marking process.

Discussion

The findings suggest a number of tentative implications of the research. First, along with the research of Sanderson and others, the current findings emphasise the importance of the process of reading and constructing a full meaning of the student's response as a part of the marking process. The codes 'reads' and 'summarises/paraphrases' were among the most frequently applicable codes and the frequency of reading behaviours seemed to be associated with marker agreement. As well as leading to the unsurprising conclusion that careful reading of responses is important to accurate marking, there may be implications for current moves towards on-screen marking as reading texts on-screen may be more difficult than reading from paper, particularly for longer texts (O'Hara and Sellen, 1997).

Secondly, evaluation processes were very important in the marking process as would be expected. Positive and negative evaluations were among the most commonly observed behaviours. Interestingly, despite the current culture of positive marking, there were fairly similar frequencies of positive and negative evaluations and frequent overlaps of positive and negative evaluations. This is in line with Greatorex's (2000) finding that although mark schemes are designed to positively reward performance with descriptions of performance written in positive terms, examiners' tacit knowledge, perhaps inevitably, leads them to view achievement in both positive and negative ways. Further, lower frequencies of positive evaluations appeared to be associated with severity and with lower marker agreement emphasising the importance of not overlooking positive elements of responses.

Thirdly, comparing a response with other responses seems to be advantageous to marker agreement. Comparisons are to be expected according to Laming (2004) who considers all judgements to be relative. Tversky and Kahneman (1974) suggest that subjects anchor subsequent judgements to initial ones. Indeed, Spear (1997) found that good work was assessed more favourably following weaker material and that high quality work was evaluated more severely following high quality work. Although assessment in UK national examinations usually aspires towards criterion-referenced standards (Baird, Cresswell & Newton, 2000) with the intention that student work is judged against criteria rather than measured by how it compares to the work of others, the findings support the view that it is necessary to have experience with a range of student work in order to understand the criteria fully and to make judgements fairly. Indeed, the importance of using exemplars in the definition and maintenance of standards is generally acknowledged (Wolf, 1995; Sadler, 1987).

The findings of this research support the view that assessment involves processes of actively constructing meaning from texts as well as involving cognitive processes. The idea of examining as a practice that occurs within a social framework is supported by the evidence of some

social, personal and affective responses. Aspects of markers' social histories as examiners and teachers were evident in the comparisons that they made and perhaps more implicitly in their evaluations. The overlap of these findings with aspects of various previous findings (e.g. the marking strategies identified by Greatorex and Suto, 2006) helps to validate both current and previous research, thus aiding the continued development of an improved understanding of the judgement processes involved in marking.

References

Baird, J., Cresswell, M. & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, **15**, 2, 213–229.

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

Greatorex, J. (2000). *Is the glass half full or half empty? What examiners really think of candidates' achievement*. A paper presented at the British Educational Research Association Annual Conference, Cardiff, available at: http://www.leeds.ac.uk/educol/documents/00001537.doc (accessed 9 January 2007).

Greatorex, J. & Suto, W. M. I. (2006). *An empirical exploration of human judgement in the marking of school examinations*. A paper presented at the International Association for Educational Assessment Conference, Singapore, 2006.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

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

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, **19**, 246–276.

Milanovic, M., Saville, N. & Shuhong, S. (1996). A study of the decision making behaviour of composition-markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment*. Cambridge: Cambridge University Press.

O'Hara, K. & Sellen, A. (1997). A comparison of reading paper and online documents. In S. Pemberton (Ed.), *Proceedings of the conference on human factors in computing systems*. 335–342. New York: Association for Computing Machinery.

Pula, J. J. & Huot, B. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, **13**, 2, 191–209.

Sanderson, P. J. (2001). *Language and differentiation in Examining at A Level. PhD Thesis*. Unpublished doctoral dissertation, University of Leeds, Leeds.

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research*, **39**, 2, 229–233.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, **185**, 1124–1131.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J.: Ablex Publishing Corporation.

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

Quality control of examination marking

John F. Bell, Tom Bramley, Mark J. A. Claessen and Nicholas Raikes Research Division

Abstract

As markers trade their pens for computers, new opportunities for monitoring and controlling marking quality are created. Item-level marks may be collected and analysed throughout marking. The results can be used to alert marking supervisors to possible quality issues earlier than is currently possible, enabling investigations and interventions to be made in a more timely and efficient way. Such a quality control system requires a mathematical model that is robust enough to provide useful information with initially relatively sparse data, yet simple enough to be easily understood, easily implemented in software and computationally efficient – this last is important given the very large numbers of candidates assessed by Cambridge Assessment and the need for rapid analysis during marking. In the present article we describe the models we have considered and give the results of an investigation into their utility using simulated data.

This article is based on a paper presented at the 32nd Annual Conference of the International Association for Educational Assessment (IAEA), held in Singapore in May 2006 (Bell, Bramley, Claessen and Raikes, 2006).

Introduction

New technologies are facilitating new ways of working with examination scripts. Paper scripts can be scanned and the images transmitted via a secure Internet connection to markers working on a computer at home. Once this move from paper to digital scripts has been made, marking procedures with the following features can be more easily implemented:

- Random allocation: each marker marks a random sample of candidates.
- Item-level marking: scripts are split by item – or by groups of related items – for independent marking by different markers.
- Near-live analysis of item-level marks: item marks can be automatically collected and collated centrally for analysis as marking proceeds.

Features such as these open the possibility of analysing item marks during marking and identifying patterns that might indicate aberrant marking. Our aim is to speed up the detection of aberrant marking by directing marking supervisors' attention to the marking most likely to be