Do assessors pay attention to appropriate features of student work when making assessment judgements?

Victoria Crisp Research Division

Introduction

This article draws on a study of the cognitive and socially-influenced processes involved in marking (Crisp, 2007; Crisp, *in press*; Crisp, *in submission*) and grading (analysis ongoing) A-level geography examinations and pilot research into the marking of GCSE coursework by teachers. These data were used to investigate the features of student work that examiners and teachers pay attention to and whether these features are always appropriate.

Where assessments involve constructed responses, essays or extended projects, the human judgement processes involved in assessing work are central to achieving reliable and valid assessment. Consequently, we need to know that appropriate features of student work influence assessment decisions and that irrelevant features do not.

Lumley (2002) suggests that less typical responses that are not accommodated in the assessment guidance force assessors to develop their own judgement strategies and they may be influenced by their intuitive impressions. If this is the case, there is the potential for criteria that are not intended to be used in marking to have an influence.

Several studies (Milanovic, Saville and Shuhong, 1996; Vaughan, 1991) have investigated marking processes in the context of English as a second language and key criteria used during assessment could be identified. Vaughan also found that different assessors (making holistic ratings) focus on different aspects of essays to each other and may have individual approaches to reading essays. Elander and Hardman (2002), in the context of psychology examinations, found that different examiners valued different factors more or less and that different factors were more predictive of the overall mark with different markers.

In the context of grading (or awarding) decisions, Cresswell (1997) found little evidence in awarders' verbalisations in meetings of how particular features of candidate work influenced decisions. Work by Murphy *et al*. (1995) found that awarders' individual views of what constitutes grade worthiness were more important in determining their decision making than other information such as statistics (although other information played a part). Further to this, Scharaschkin and Baird (2000) found that the degree of consistency of student work within a script, a feature that was not a part of the mark scheme guidance, influenced grading decisions for biology and sociology A-level scripts.

Sanderson (2001) developed a model of the process of marking A-level essays which emphasised (amongst other things) the social context of assessment judgements. Cresswell (1997) identified affective reactions to scripts (e.g. like or dislike) by examiners in awarding meetings. It is hypothesised that social, personal and affective reactions could perhaps affect the features attended to by assessors and explain some differences between examiners in terms of marks awarded.

The main focus of the research studies drawn on here was to improve our understanding of the judgement processes involved in marking and grading by examiners and marking by teachers. However, the focus of the additional analyses for this paper was on investigating whether assessors pay attention to appropriate features of student work when making assessment judgements.

Method

This article draws on data from two research studies both using verbal protocol analysis methodology. Verbal protocol analysis involves asking participants to complete a task whilst 'thinking aloud' and then using the verbalisations to infer the processes going on. This is generally considered a suitable method for investigating cognitive processes but has limitations in that certain types of information or processes do not occur at a conscious level and so can not be reported by participants (Ericsson and Simon, 1993).

The first set of data drawn on in this paper was collected in the context of A-level geography examinations and the main analyses have been reported in Crisp (2007; *in press*; *in submission*). Six experienced examiners were involved in the research and after some initial marking each examiner marked four to six scripts from each exam whilst thinking aloud. Each examiner also carried out a grading exercise for each exam whilst thinking aloud in which they were asked to judge the A/B boundary for the paper (i.e. to judge the minimum mark worthy of an A grade). During the grading exercise examiners had access to relevant parts of the Principal Examiner's report to the awarding team and had two scripts on each of the marks within the range used in the original awarding meeting. The grading exercises aimed to simulate and gain insight into the cognitive aspects of grading judgements without interference from the potential influence of social or political dynamics of live awarding meetings.

The second set of data drawn on in this paper was collected for pilot research in the context of GCSE coursework. One English teacher and one Information and Communications Technology (ICT) teacher each marked two coursework pieces at home and then later marked two further pieces whilst thinking aloud.

With both these sets of data the verbal protocols were analysed in detail using appropriate coding schemes (see, for example, Crisp, *in press*). A range of types of assessor behaviours and reactions were identified including reading behaviours, evaluations and personal, affective and social reactions.

With the A-level data the frequencies of different types of behaviours were compared between the exams and between examiners (see Crisp, 2007; Crisp, *in press*). Tentative models of the marking process and the grading process were developed by investigating patterns of behaviours/codes and the likely cognitive processes were considered in relation to existing theories of judgement (Crisp, *in submission*). This work

identified that evaluations either occurred alongside reading ('concurrent evaluations') and involved an evaluation of a part of the work, or occurred at a more overall level ('overall evaluations') and involved bringing together the understanding of the student's response, including its strengths and weaknesses, and beginning to convert this to a mark or grade decision (Crisp, *in submission*).

With the data from GCSE coursework marking, the teacher behaviours and reactions were compared between subjects (though with some caution given that there was only one teacher in each subject in this pilot work).

Results

For this article, additional analyses of the data were conducted. This involved reviewing extracts of the verbal protocol transcripts where assessors paid attention to particular features of student work or showed particular reactions, and then ascertaining whether these features affected evaluations. Evaluations were found to occur either concurrently with reading (usually an evaluation of a particular element of the student work) or after reading is complete as part of an overall evaluation and consideration of the appropriate mark. This distinction will be used to structure the analysis. This article focuses mostly on the data from A-level geography marking. It will consider data from the A-level geography grading exercises and the GCSE coursework marking pilot research more briefly.

Geography A-level marking and grading

Most aspects noted by examiners were closely related to the mark scheme and were about geography content knowledge, understanding and skills. Additionally, examiners sometimes made comments relating to aspects of students' attempts to achieve the requirements of the task ('task realisation') (see Crisp, *in press*). These included comments on the length of a response, noting whether the student had understood the question, commenting on the relevance of points and on material missing from a student's response (Crisp, 2007; Crisp, *in press*). Most of the features noted by examiners in this category are likely to be legitimate influences on examiner judgements. One exception might be the length of responses which probably should not affect marks directly. A further more detailed look at the verbalisations coded in this category revealed that all evaluative comments on length related to the response being shorter than expected and hence not showing sufficient knowledge, understanding and skills, or being longer than expected and including too much information that is not necessarily used to directly answer the question. In both cases it then becomes acceptable for these factors to affect examiner judgements as they are aligned with the marking criteria.

References to the geography A-level Assessment Objectives during marking were coded in the analysis (Crisp, 2007; Crisp, *in press*) as this gives insight into how examiners convert what they have seen (possibly categorising and combining cues or information) into marks. The high frequency of reference to Assessment Objectives (6.88 references to an Assessment Objective per script on average during marking) and the fairly frequent association with positive or negative evaluations (5.97 instances on average per script of a reference to an Assessment Objective co-occurring with a positive or negative evaluation) gives a strong indication that markers do tie their thinking closely to the valued aspects of the mark scheme guidance (i.e. the intended marking criteria). There was also fairly frequent reference to the mark scheme during marking (2.03 times on average per script). The analysis will now focus on aspects of marker verbalisations that were less expected and less clearly related to the qualities described in the mark scheme.

Language

Examiners sometimes commented on the quality of a student's language use or on orthography (i.e. handwriting, legibility and presentation) (see Crisp, 2007; Crisp, *in press*). This occurred 1.46 times per script on average during marking. A more detailed analysis of the marking transcripts for each of the 86 instances revealed that 27 instances were not associated with any evaluation, 58 instances were associated with either a positive or negative concurrent evaluation (i.e. an immediate evaluation made during the process of reading the response), 24 instances fed into overall evaluations relating to Communication as an Assessment Objective, and 10 instances were associated with overall evaluations that were not specifically linked to assigning marks for communication[1].

This suggests that language quality rarely impacts on overall evaluations except where communication is an explicit criterion for evaluation (as in the A2 exam). Instances where reference to language use did feed into overall evaluations occurred where the structure was weak resulting in a reduced clarity in the student's meaning or where the legibility of the response was sufficiently weak to impair understanding of the student's meaning and line of argument. It seems that language only affects overall evaluations where communication is an aspect intended to be assessed or in circumstances where the quality of language or handwriting impairs understanding.

It is interesting that in a number of the instances where language quality or orthography was associated with a concurrent evaluation examiners said that a response would get a certain number of marks *despite* its weak structure or expression. This might suggest that they are in control of the influences on their marking and prevent language skills from impacting their judgements where marking guidance determines that it should not.

Of the 28 instances of reference to language use during grading, 22 were associated with a concurrent evaluation (e.g. 'sound introduction, quite well written') and 7 were associated with the overall evaluation of the quality of the script. In the instances that fed into overall evaluations it seems that language quality was occasionally one factor in the examiner's mind when attempting to make a judgement of grade worthiness even when it was not an explicit mark scheme criterion. However, it is interesting to note that all comments on language which seemed to feed into overall evaluations were positive rather than negative.

Social perceptions

As noted in Crisp (Crisp, *in press*) examiners sometimes appear to have social perceptions of students during marking as understood from characteristics of the script. Markers sometimes made assumptions about other characteristics of students (0.85 per script on average) or inferred likely further performance of the student (0.39 per script on average).

The code 'assumptions about candidates' was applied where an examiner inferred student characteristics (e.g. ability, lazy, thoughtful)

1 In this and the analyses that follow some instances of a particular code were associated with both a concurrent and an overall evaluation. Consequently the numbers quoted sometimes add up to more than the total number of instances.

or inferred how a student has approached the task from the student's response. Reviewing transcript extracts revealed that assumptions about candidates were often about general geography ability or specific aspects of knowledge (e.g. knowledge of place) and were hence part of the examiner's progress towards forming an overall impression of a student's relevant abilities. Detailed analysis of the 50 instances of this code found that 17 instances were not associated with an evaluation, 26 instances were associated with a positive or negative concurrent evaluation, and 26 instances were issues that fed into overall evaluations and so may have influenced the marks awarded. Of the 26 instances of assumptions about candidates being linked to overall evaluations 23 were at least partly about the student's geography ability or knowledge, for example: '*this lad knows a lot, likes to write a lot*'. The three instances linked to overall evaluations that did not relate to geography ability still related closely to the students' attempts to answer the questions.

In grading, assumptions about candidates were infrequent (0.13 times per script on average or 12 instances in total). In a similar way to during marking, instances sometimes related to concurrent evaluations (5 instances) or overall evaluations (3 instances) but were usually assumptions relating to geography abilities or to do with the students' attempts to answer the questions. As with marking, such assumptions seem to aid the examiner in synthesising their understanding of different aspects of the student's response in order to come to an understanding of the overall level of performance.

Examiners occasionally made predictions about candidate performance before finishing reading a response or sometimes even before beginning to read (Crisp, 2007; Crisp, *in press*). Predictions related to the likely quality of the response or to the kinds of material they expected to see in the rest of the response or script, for example: '*This is not going to be a better paper, is it?*'

Analysis of the 23 instances of performance predictions (from the marking protocols) found that 7 involved no evaluation, 16 included a concurrent evaluation (e.g. '*not going to be a strong script I think*') and 5 were associated with considering the overall performance. Where predictions are associated with the overall evaluations these often occurred later in the reading of a response (when the examiner has more information and so it is more reasonable for them to make an overall prediction). The rest of the response was still read carefully and the entire view of the script was checked against the marking criteria.

There were very few instances of examiners predicting performance in the grading data (0.04 per script on average) and these were similar in nature to the instances during marking (expecting certain content, hoping response will get better). Only 1 of the 4 instances contained an evaluation in grading and this was a concurrent rather than an overall evaluation.

Personal and affective reactions

Examiners sometimes showed affective (i.e. emotional) or personal reactions to features of students' work (Crisp, 2007; Crisp, *in press*). During marking, positive affect (e.g. '*so good he is on target now, I'm really pleased*') was shown 0.75 times per script on average and negative affect was displayed 1.24 times per script on average. Examiners showed amusement or laughed during marking 0.49 times per script on average and showed frustration 0.39 times per script on average.

There were a total of 44 instances in total of examiners showing positive affect (or sympathy) towards students and/or their work during marking. Of these, 20 instances were not associated with an evaluation, 20 were linked to a concurrent evaluation and 5 were linked to an overall evaluation. Instances of positive affect being linked to concurrent evaluations usually involved a positive feature of a script eliciting both a positive evaluation and positive affect (e.g. '*oh hooray, hooray, hooray, someone has actually thought about that!*') or a feature of the script eliciting sympathetic feelings and a negative evaluation. In both types of instances it is the positive or negative evaluation and not the examiner's affective reaction which may be going on to influence further evaluation.

In grading, evidence of positive affect was fairly infrequent and the verbalisations showing positive affect were similar in nature to those occurring during marking.

There were 73 instances of examiners showing a negative affective reaction to student work (e.g. '*oh no not the flippin' Italian dam again*') during marking. Of the instances, 41 were not associated with any evaluation, 27 were associated with a concurrent evaluation and 6 were associated with an overall evaluation. Looking at the instances of links with concurrent and overall evaluations suggests that, similarly to positive affect, negative affect is usually a response to negative aspects of students' responses in terms of the knowledge and skills required, or a response to efforts to appropriately answer questions. Some verbalisations also indicated that examiners were sufficiently aware of their emotional responses to not allow these to influence the marks they award. Negative affective reactions were infrequent in grading. Most instances were not associated with evaluations and those that were, were similar in nature to the instances in marking.

In marking, there were 29 instances of laughter or amusement in response to student work. Only 6 instances were linked to concurrent evaluations and none to overall evaluations. The concurrent evaluations tended to occur where a student gave certain kinds of factually incorrect information which are then evaluated as incorrect. Amusement and laughter were infrequent in grading and were only associated with a concurrent evaluation on one occasion.

Frustration or disappointment was shown by examiners in 23 instances in relation to marking. In 7 instances this was not connected to evaluations, in 13 it was linked to a concurrent evaluation and in 4 instances to an overall evaluation. Where examiners showed frustration or disappointment linked to a concurrent or overall evaluation this tended to be where the student's work was weak in some respect, something was missing from their response or their response was not appropriately targeted to the question. In grading frustration was infrequent. As with marking more than half of these instances were related to some kind of evaluation but they appeared to relate to legitimate weaknesses in student work.

It seems that although a number of different types of emotive reactions were elicited from examiners, these affective responses were caused by qualities of the geography or students' abilities to achieve the task, and it was this rather than any emotional response that guided marking and grading decisions.

GCSE coursework marking

This section will describe briefly the features attended to by teachers when marking GCSE coursework using the pilot study. These data do need to be treated with some caution due to the small scale of this pilot work but may provide insight into whether the findings in A-level geography are likely to generalise to marking by teachers, marking in other subject areas and marking of a different type of student work.

First, it is worth noting that the teachers referred to the marking guidance fairly frequently, and particularly frequently in ICT (19.5 times per coursework piece for ICT and 3.5 times per coursework folder in English on average). The difference in frequency between subjects relates to the nature of the mark schemes. The ICT mark scheme includes very specific task elements that students need to show in their work, and hence requires very close reference to the mark scheme during marking. The mark scheme for the English coursework represented a continuum on a number of different types of skills and thus appears to be easier for teachers to internalise, such that they do not need to refer to it as frequently.

In the pilot work it was considered useful to code the detailed features of student work commented on by teachers in their verbalisations to allow investigation of differences between subjects. In English these included:

- evaluates spelling, punctuation or grammar
- evaluates style, vocabulary, quality of expression, use of technical terminology or text structure
- evaluates imagination, sophistication, whether interesting or formulaic
- student's personal response to literary texts
- making comparative points about texts/poems
- understanding of genre
- student's use of quotations from literature
- presence of/quality of conclusions to essays
- use of narrative

In ICT features focussed on included:

- evaluates spelling, punctuation or grammar
- evaluates style, vocabulary, quality of expression, use of technical terminology or text structure
- use of IT and non-IT source materials
- absence/presence of information or evidence on the sources used
- designs/image editing
- saving files and folders
- use of number
- spell-checking and proof-reading

These are all features included in the relevant marking criteria and are hence intended and legitimate influences on marking decisions.

Again there were other behaviours (either features of the work being noted or reactions occurring in response to features of the work) apparent in the transcripts which are less obviously related to intended influences on marking. These were similar to those seen in A-level exam marking and included:

- commenting on orthography
- commenting on aspects of task realisation (e.g. response length)
- affective reactions and amusement
- social perceptions (e.g. predicting performance, reflections on characteristics of students)

Looking at the verbalisations fitting these codes suggests that, similarly to the marking and grading of A-level geography, inappropriate features of student work do not appear to influence evaluations in ways that they should not.

Discussion

The verbal protocol methodology was generally a successful method for exploring the features of student work attended to during marking. However, the limitation of the method in terms of verbal protocols not supplying a complete record of all thoughts passing through working memory (Ericsson and Simon, 1993) is problematic. Therefore, we cannot be completely sure that no inappropriate features of student work ever influenced overall evaluations and mark decisions in unintentional ways although the data are encouraging in this respect.

The data collected suggest that assessors mostly attend to features of student work related to intended marking criteria during their marking or grading process and that they focus mostly on the intended marking criteria in their actual evaluations. Most of the verbalisations focussed on features relevant to the subject knowledge, understanding or skills under assessment and Assessment Objectives and the marking guidance were used fairly frequently. There were, however, some types of behaviours or reactions during their processing that might, at first inspection, indicate that assessors sometimes attend to features of student work that are not within the intended focus of evaluations. Analysis of these instances revealed that where features were attended to that were not indicated by the mark scheme these did sometimes influence ongoing evaluations and occasionally fed into overall evaluation and mark consideration. However, close analysis indicated that most instances were actually caused by features of the student work that were intended to be evaluated. Additionally, several verbalisations indicated that although features were noted and sometimes considered during evaluations, assessors tended to be in control of whether these influenced actual marks.

Given that inappropriate features of student work and personal, social and affective reactions did not appear to influence overall evaluations and mark consideration inappropriately, it seems that such behaviours do not explain variations in marks between examiners. This may suggest that variations are a result of other factors perhaps such as variations in the weight that examiners place on different features, variations in the extent to which examiners are willing to be lenient when inferring a student's knowledge behind a partially ambiguous response, or variations in the interpretation of aspects of the mark scheme. These issues would require further investigation to ascertain their contribution.

The data are consistent with the view that the judgement processes involved in the assessments investigated rely closely on professional knowledge and that evaluations of work are strongly tied to values communicated by the mark scheme. Features relating to task realisation also legitimately influence evaluations. Thoughts regarding language use, social perceptions and affective reactions also sometimes led to concurrent evaluations and occasionally fed into overall evaluations but assessors were in control of influences on their judgements and no inappropriate biases were found using the current methods.

Note:

This article is based on a paper presented at the International Association for Educational Assessment Annual Conference in Baku, Azerbaijan, September 2007.

References

Cresswell, M. J. (1997). *Examining judgements: theory and practice of awarding public examination grades*. PhD Thesis. Unpublished doctoral dissertation, University of London, Institute of Education, London.

Crisp, V. (2007). *Comparing the decision-making processes involved in marking between examiners and between different types of examination questions*. Paper presented at the British Educational Research Association Annual Conference, London.

Crisp, V. (*in press*). Exploring the nature of examiner thinking during the process of examination marking, *Cambridge Journal of Education*.

Crisp, V. (*in submission*). Towards a model of the judgement processes involved in examination marking.

Elander, J. & Hardman, D. (2002). An application of judgment analysis to examination marking in psychology. *British Journal of Psychology*, **93**, 303–328.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: verbal reports as data*. London: MIT Press.

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.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. & Gower, R. (1995). *The dynamics of GCSE awarding*. Report of a project conducted for the School Curriculum and Assessment Authority, School of Education, University of Nottingham.

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

Scharaschkin, A. & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal*, **26**, 3, 343–357.

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.

Marking essays on screen: towards an understanding of examiner assessment behaviour

Stuart Shaw CIE Research

Introduction

Computer assisted assessment offers many benefits over traditional paper methods. In translating from one medium to another, however, it is crucial to ascertain the extent to which the new medium may alter the nature of the assessment and marking reliability. Appropriate validation studies must be conducted before a new approach can be implemented in high stakes contexts. The pilot described here is the first attempt by Cambridge International Examinations (CIE) to mark, on-screen, extended stretches of written text for the Cambridge Checkpoint English Examination. The pilot attempts to investigate marker reliability, construct validity and whether factors such as annotation and navigation differentially influence marker performance across the on-paper and on-screen marking modes.

Candidates wrote their answers on paper scripts in the normal way. The scripts were then scanned and digital images of them were sent by secure electronic link to examiners for on-screen marking at home using Scoris® software.

It can be relatively hard for examiners to make a full range of annotations when marking on screen. For this reason annotation sophistication was manipulated in the pilot as well as marking mode. Four marking methods were compared: on-paper with sophisticated annotations (current practice), on-paper with simplified annotations, on-screen with sophisticated annotations, and on-screen with simplified annotations.

The research literature

There is a large research literature relevant to this project. Key aspects of this literature are summarised below.

Comparability of marking across on-screen and on-paper modes

The literature is mixed on this topic.

- Bennett (2003) carried out an extensive review of the literature and concluded that 'the available research suggests little, if any, effect for computer versus paper display' (p.15).
- Differences were found in a few studies not reviewed by Bennett, however, e.g. Whetton and Newton (2002) and Royal-Dawson (2003).
- Sturman and Kispal (2003) observed quantitative differences between online and conventional marking of tests of reading, writing and spelling for pupils typically aged 7 to 10 years, but an analysis of mean scores showed no consistent trend in scripts receiving lower or higher scores in the e-marking or paper marking: 'absence of a trend suggests simply that different issues of marker judgement arise in particular aspects of e-marking and conventional marking, but that this will not advantage or disadvantage pupils in a consistent way' (p.17). Sturman and Kispal concluded that e-marking is at least as accurate as conventional marking. Wherever differences between the