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

Victoria Crisp
Cambridge Assessment

A paper presented at the International Association for Educational Assessment Annual Conference, Baku, Azerbaijan, September 2007.

Victoria Crisp
Core Research Group
Research Division
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU
UK
Direct dial. +44 (0)1223 553805
Fax. +44 (0)1223 552700
Email: crisp.v@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge.

Cambridge Assessment is a not-for-profit organisation.

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

IAEA conference sub-theme: Evaluating the quality of assessment

Abstract

It is via the judgements of appropriate experts that assessment decisions are made yet the actual thought processes involved during marking or grading are under-researched. This paper will draw on a study of the cognitive and socially-influenced processes involved in marking and grading A level geography examinations and pilot research into the marking of GCSE coursework by teachers. These data will be used to investigate whether assessors pay attention to appropriate features of student work.

Verbal protocols of assessors' thinking aloud whilst marking and grading work were collected and measures of marker agreement were obtained. The protocols were analysed in detail using appropriate coding schemes. From the behaviours identified, a tentative model of the marking process was developed, within which features of student work affecting judgements and social and personal reactions were identified. Whilst many features that appeared to influence evaluations were clearly focussed on the criteria intended for evaluation, some were not and could have influenced evaluations. Reactions to language use or legibility (when not assessing communication), personal or emotional responses and social responses sometimes occurred before marking decisions. The paper will discuss whether such responses could explain variations in marks from different examiners.

This paper draws on research data reported elsewhere and work still in progress and expands on some of the analyses previously conducted. The papers listed below report on different aspects of the data analysis linked to research involving A level geography examination marking.

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 submission (a)) Exploring the nature of examiner thinking during the process of examination marking, *Cambridge Journal of Education.*

Crisp, V. (to be submitted) A tentative model of the judgement processes involved in examination marking.

Introduction

Where assessments involve constructed responses, essays or extended projects, the human judgement processes[1] involved in assessing work are central to achieving reliable and valid assessment yet they are not fully understood. As assessments often have significant outcomes for students, understanding the decision-making processes underlying marking and grading judgements is important. Additionally, we need to know that the appropriate features of student work influence assessment decisions and that irrelevant criteria do not. This paper looks at these issues in particular by drawing on data from research presented elsewhere. These research studies involved A level[2] geography examiners 'thinking aloud' whilst marking scripts from two different exams (Crisp, 2007; Crisp, in submission (a); Crisp, to be submitted) and 'thinking aloud' whilst carrying out a grading exercise (analysis ongoing). The discussion will also draw on pilot research in which an English teacher and an Information and Communications Technology (ICT) teacher 'thought aloud' whilst marking some pieces of GCSE[3] coursework. The research question that this paper will try to answer is whether assessors pay attention to appropriate features of student work or whether inappropriate features are sometimes attended to. It will also discuss whether any inappropriate features attended to appear to go on to influence the actual awarding of marks or grading judgements. These issues are important to the quality of assessment. A further aim was to investigate whether verbal protocol analysis could facilitate investigation of these issues.

[1] This research is concerned with marking and assessment as conducted by expert human assessors and does not address the parallel assessment issues that may exist where work is marked electronically.

[2] A levels are national general qualifications taken by many students in the UK at age 17/18 years and A level results are often a substantial factor in university entrance decisions. For any one A level course in a particular subject students take a number of units some at AS (Advanced Subsidiary) level and some at A2. AS units, which are less demanding, are taken earlier and can be used to gain an AS qualification. A2 units, which are more demanding, are taken later to achieve a full A level qualification. The units are assessed individually often via a traditional pen-and-paper exam. A levels are also available internationally via some UK-based Awarding Bodies.

[3] GCSEs are national general qualifications taken by many pupils in the UK at age 16 years. GCSE results are often a requirement for progression to A level. For each GCSE subject pupils sit one or more exams and are sometimes required to submit a piece of extended work (coursework) which is marked by teachers and then externally moderated. International GCSEs are available to students outside the UK.

Sanderson (2001) and others (e.g. Broadfoot, 1996; Filer, 2000; Gipps, 1999; Shay, 2004; Shay, 2005) have emphasised the socially-framed nature of assessment decisions. Examiners are part of examining teams which can be considered communities of practice (Sanderson, 2001). Members of communities of practice form shared understandings and marking practices through their interactions (Wenger, 1998). These experiences, along with their personal professional knowledge and experience, will determine the influences on their assessment judgements. Teachers are part of a community of teaching professionals in their subject and a department within a school could be considered a community of practice (Boaler, 1999). These small communities of practice may be linked to a wider, less closely connected community of practice involving other schools. Whilst experience and shared understandings from ‘community’ interactions amongst teachers will encompass many areas of thought and experience other than just assessment, experiences are likely to impact on their marking judgement processes.

An important first step to making assessment judgements is reading a response and forming a mental representation of that response that corresponds closely to the student’s intended meaning. This may seem like the simplest aspect of the cognitive processes involved in marking, and it may well be unproblematic in most cases, but theories of reading (e.g. Gernsbacher, 1990; van Dijk & Kintsch, 1983) tend to see reading as an active cognitive process in which interpretation and inference are used to build a representation drawing on pre-existing knowledge structures. This means that there is space for different interpretations of a student’s meaning.

General theories of judgement may be relevant to the current investigation. Some models of judgement (termed ‘analytic’ by Sadler, 1989) such as that described by Einhorn (2000), suggest that a judge initially identifies cues or criteria, measures these cues in some way, and then combines these measurements by aggregation and applying weightings (either via a rigid formula or more flexibly). Other models (termed ‘configurational’ by Sadler,

1989) suggest that the judgement of quality is produced immediately and directly in response to the material and that criteria are used subsequently to explain the judgement. In terms of the criteria used in assessment, Sadler (1989) proposes that there is a wide pool of possible criteria that can be considered during an assessment but that only a sample of these will be used at any time. Sadler uses the terms ‘manifest’ and ‘latent’ to describe criteria that are actively in use (i.e. manifest criteria that the marker intends to use) and criteria that can be triggered in response to something in the student work (i.e. latent criteria) and can become part of the working set of manifest criteria. Whether criteria or cues are identified before or after overall judgements will be interesting to see in the data. The issue of whether there is a pool of possible criteria with some in use and others available if needed will also be of relevance to the investigation of whether appropriate features of student work are attended to by assessors. Some reasoning and judgement processes are thought to occur at an unconscious level (e.g. dual processing theories, see Sloman, 2002; theories of conscious and unconscious competence, see Dreyfus & Dreyfus, 2005). Consequently, it is possible that some criteria are used during marking at an unconscious rather than a conscious level. These may be difficult to capture with the think aloud methodology.

Kahneman and colleagues (e.g. Gilovich, Griffin, & Kahneman, 2002) have investigated and theorised a number of cognitive heuristics (or shortcuts) that can come into use in judgements. These are often efficient strategies supporting sufficiently accurate judgements but can also lead to unintentional biases. The availability heuristic is most relevant to the use of appropriate cues or information when making judgements. The availability heuristic may be triggered if the information required is not available. In such cases available information that is linked but not entirely appropriate is substituted for the required information when the judgement is made. In this way, if a certain type of information is more easily available than the intended criteria, this may influence decisions.

A further recurring issue in marking judgement research relating to cognitive processing is the relationship between impressionistic views and marking criteria. Several researchers (e.g. Vaughan, 1991; Huot, 1993; Lumley, 2002; Delaney, 2005) have commented on the sometimes problematic relationship between the marking guidance, the examiner's impression of the response, and the response itself. Lumley (2002) argues that the assessor's task is "to reconcile their impression of the text, the specific features of the text, and the wordings of the rating scale, thereby producing a set of scores." (p. 246). He suggests that less typical responses that are not accommodated in the assessment guidance force assessors to develop their own judgement strategies. Lumley suggests that "in doing this they try to remain close to the scale, but are also heavily influenced by the complex intuitive impression of the text obtained when they first read it" (p. 246). If this is the case it means that there is the potential for criteria that are not formally intended to be used in marking to have an influence.

Several studies (Milanovic, Saville, & 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. Greatorex and Suto (2006) identified cognitive marking strategies in the context of GCSE marking and found some variation between examiners although this did not seem to affect marking reliability. 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. However, these findings do not tell us whether the features attended to were appropriate or whether the differences have a negative impact on marking consistency.

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. Cresswell went on to argue that direct overall evaluations are made and revised via "an evolutionary succession of direct evaluations" (p.289). 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). Either of these findings might be considered potentially problematic. 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. Others (e.g. Delaney, 2005) have noted how assessors sometimes enter into a constructed dialogue with students when judging work. Cresswell (1997) identified affective reactions to scripts (e.g. like or dislike) by examiners in awarding meetings (meetings at which grade boundaries are decided). 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.

It is via the judgements of appropriate experts that assessment decisions are made yet the actual thought processes involved during marking and grading are under-researched. Understanding such processes may have beneficial implications for assessor training, for technological changes to assessment systems and for ensuring assessments are valid and reliable.

The main focus of the research which is drawn on here was to improve our understanding of the decision-making processes involved in marking and grading by examiners and marking by teachers and to compare the processes involved in marking shorter questions and essays. However, the particular focus of this paper is to use these data to investigate 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. In verbal protocol analysis (e.g. Ericsson & Simon, 1993) participants are asked to articulate their thoughts as much as possible whilst conducting a task of interest. The verbal protocols are analysed and used to infer the underlying processes involved in the task. There have been some criticisms of verbal protocol analysis (e.g. Nisbett & Wilson, 1977) but it is generally considered a useful method if used appropriately (Ericsson & Simon, 1993). Previously, verbal protocols have provided findings relating to whether marking schemes are used to the full (Sanderson, 2001) suggesting that it is likely to be a useful method for the current research. However, there are acknowledged limitations to the method, perhaps most importantly that certain types of information or processes do not occur at a conscious level and so can not be reported by participants (Ericsson & Simon, 1993). Consequently, verbal protocols of examiners marking may not capture all features attended to and all influences on mark decisions.

The first set of data drawn on in this paper was collected in the context of A level[4] geography examinations. Aspects of this work are reported in Crisp (2007; in submission (a); to be submitted). In the research two contrasting examination papers were chosen: one AS unit involving short to medium length responses and one A2 level unit requiring students to write two essays from a choice. 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).

[4] For any one A level qualification in a particular subject students take a number of units some at AS (Advanced Subsidiary) level and some at A2. AS units, which are less demanding, are taken earlier and can be used to gain an AS qualification. A2 units, which are more demanding, are taken later to achieve a full A level qualification. The units are assessed individually often via traditional pen-and-paper examinations.

After examination scripts have been marked a team of senior examiners meet and with reference to a range of information (e.g. live scripts, archive scripts from previous years, statistical information, Principal Examiner's[5] reports) judgementally decide the minimum mark worthy of key grades. For each judgementally determined boundary, the meetings often involve the examiners looking at individual scripts on a range of marks expected to contain the boundary mark and deciding whether each script is worthy of grade A or not. Looking at a range of scripts and discussing their thoughts on grade-worthiness leads to a group decision on the minimum mark for the boundary. For a detailed description of the awarding related issues see, for example: QCA Code of Practice (2007), Cresswell (1997), French et al. (1992). For the grading exercise in the research drawn on here 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 recommended by the Principal Examiner as containing the A grade boundary. 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 examiners were interviewed about their marking and grading judgement processes for each exam.

The second set of data drawn on in this paper was collected for pilot research in the context of GCSE coursework in English and Information and Communication Technology (ICT). The written coursework element for English was selected along with one coursework element of an ICT GCSE. One English teacher and one ICT teacher participated. Each marked two coursework pieces (or two coursework folders) at home and then later marked two further pieces whilst thinking aloud. The teachers were interviewed about their marking process. These data are referenced in this paper more briefly than the geography A level exam marking data due to their small scale nature.

[5] A Principal Examiner for an examination is responsible for writing the examination questions (with review and input from a question paper committee) and for coordinating the marking of that paper as well as being involved in awarding decisions. Before an awarding meeting, each Principal Examiner prepares a report on how the examination they are responsible for performed, including a recommendation for the range of marks within which key grade boundaries are likely to lie.

However, it is useful to consider them here as they may provide tentative indications of whether the other findings are likely to generalise beyond A level geography.

Results

In both research studies assessor behaviours and reactions were identified and appropriate coding frames were developed to accommodate these. With the A level geography study, one coding frame was developed for the marking protocols (see Crisp, 2007; Crisp, in submission (a)) and this was later adapted for the grading protocols (analysis ongoing). A different coding frame was found to be more appropriate for the coursework marking protocols but the general nature of the groups of codes was similar. All protocols were coded using the coding frames.

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 submission (a)). 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, to be submitted).

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). The patterns of behaviours/codes were also used to make sense of the overall processing.

In the context of A level geography marking, analysis of the sequences of the coded behaviours apparent in the verbalisations allowed a tentative model of the marking process to be constructed. This is described in Crisp (to be submitted) but is mentioned here as it helps to frame the analysis to come. This work identified that initial evaluations of parts of a response occurred concurrently with the process of reading and building a mental representation of the student’s meaning. Concurrent evaluations are sometimes associated with Assessment Objectives (aspects that a qualification aims to assess, such as knowledge and understanding), especially when marking essays. Reading a student’s response can also trigger thoughts regarding the language used by the candidate, the student’s efforts to complete the task and social and personal reactions from examiners and these were sometimes directly associated with, or followed by, a concurrent evaluation. The research also identified that at the completion of reading the response examiners evaluated the response in a more overall way, possibly weighing up its quality, commenting on strengths and weaknesses, referring back to the mark scheme and recalling their earlier thoughts and evaluations about features such as language use and achieving the task. During this phase the examiner works towards quantifying the quality of the response with respect to the mark scheme, eventually leading to a mark decision.

The analysis below will describe the features attended to during concurrent evaluations and try to ascertain whether these features affected evaluations occurring concurrently with reading and/or fed into overall evaluations and mark consideration. This paper will focus on the data from A level geography marking to illuminate whether assessors pay attention to appropriate features when marking and the influences on assessor judgements. 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

Within the verbal protocols of marking the features or qualities of student work affecting judgements and social and personal reactions were identified. The same types of verbalisations were found in the verbal protocols of grading in terms of the types of features attended to and social and personal reactions (analysis ongoing). However, almost all behaviours occurred with much lower frequency per script in grading than in marking because any script is considered more briefly and because the marks awarded provide substantial information. Mostly the findings for grading were similar to those for marking in terms of the nature of verbalisations and whether they impacted on evaluations. Where differences were found these will be discussed.

As described in Crisp (Crisp, 2007; Crisp, in submission (a)) the codes used in analysing the protocols 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 e.g. quality of language, orthography);
- ‘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/grade’ (codes relating to assessment objectives, quantifying judgements in marking and decisions regarding grade worthiness and grade boundaries).

Note 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, to be submitted). In the analysis described below extracts of the verbal protocols were reviewed to identify whether reference to certain features or certain reactions were part of a concurrent evaluation of students’ work and/or whether there is evidence that these features or reactions were still in the assessor’s mind at the point when overall evaluations were made and hence potentially influenced decisions.

Most aspects noted by examiners were closely related to the mark scheme and were about geography content knowledge, understanding and skills. These aspects were not coded in detail as they are intended to be assessed. However, examiners sometimes made comments relating to aspects of students' attempts to achieve the requirements of the task and some of these features were captured with the codes under the title 'task realisation'. Examiners sometimes commented on the length of a response, noted whether the student had understood the question or was addressing the question, commented on the relevance of points and commented on material missing from a student's response (Crisp, 2007; Crisp, in submission (a)). 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 we would not wish to affect marks directly. Looking at the verbalisations coded in this category in more detail (occurring 0.29 times per script on average in marking and 0.22 times per script on average during grading) it is clear 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 asked. In both cases it then becomes acceptable for these factors to affect examiner judgements as they are aligned with the marking criteria.

A level (and other) qualification specifications outline the 'Assessment Objectives' (AOs) to be taught and later assessed. For this geography A level these are 'AO1 show knowledge of the specified content', 'AO2 show critical understanding of the specified content', 'AO3 apply knowledge and critical understanding to unfamiliar contexts' and 'AO4 select and use a variety of skills and techniques, including communication skills appropriate to geographical studies'. References to the geography A level Assessment Objectives during marking were coded in the analysis (Crisp, 2007; Crisp, in submission (a)) 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). These findings strongly suggest that (as we would expect) markers do give a large proportion of their attention to appropriate features of the student work during assessment judgements so we know that markers can and do use appropriate features. 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 because if features not described in the mark scheme were to affect judgements this would be a worry. Identifying any such problems would allow us to attempt to resolve them perhaps with examiner training or guidance. We may find of course that there are aspects that are attended to that do not seem appropriate at first inspection but that these do not in fact adversely impact on marking decisions.

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 submission (a)). 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[6].

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. Whilst examiners are required to check student work carefully for credit-worthy points made by candidates, it is only fair to other students that examiners do not over-infer what they think the student was trying to express in a case where their meaning is unclear. Hence, it is appropriate that these language issues noted by examiners feed into overall evaluations and mark assignment. 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. For example, one examiner said *'again it's a poorly structured answer but I fear it has put her into level 2, 5 marks'*. This would indicate instances where examiners are aware of weaknesses in language and evaluate language during their process of reading and understanding the student's meaning but 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.

Whilst wider debates continue regarding whether it is sound for students to achieve good grades in subject qualifications if their literacy skills are weak

[6] 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.

(see for example Massey & Dexter, 2002; or the Tomlinson report, Working Group on 14-19 Reform, 2004), it is desirable that examiners restrict their evaluations of language quality to instances where the mark scheme dictates assessment of communication skills (i.e. this has been decided in the specification development).

Reference to the student's language use was less frequent in grading than marking (as with most behaviours) occurring 0.30 times per script on average. Of the 28 instances 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 and consideration of overall quality of the work 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, the extent of impact or weight placed on this is hard to determine from the protocols. Communication skills are part of one of the general geography A level Assessment Objectives and hence how well the student has communicated their knowledge and understanding could be argued to be an acceptable element at the level of grade boundary decisions even where it is not an explicit marking criteria for a paper. It is interesting that all comments on language which seemed to feed into overall evaluations were positive rather than negative.

Social perceptions

Social psychology tells us that we understand new people in the social world in terms of typifications and that these influence our interactions with them (Berger & Luckman, 1967). In a similar way examiners sometimes appeared to have social perceptions of students during marking as understood from characteristics of the script. They occasionally inferred a student's gender from a characteristic of the script (0.03 times per script on average) during marking but this was never linked to a concurrent or overall evaluation. They sometimes made assumptions about other characteristics of students (0.85 per script on average), inferred likely further performance of the student (0.39

per script on average) or occasionally even made inferences about the teaching received (0.10 times per script on average). Whilst these sorts of comments from examiners do not indicate the particular features noted in the script per se, they indicate reactions that will have been caused by aspects of the student work and these could have the potential to influence evaluations.

The code ‘assumptions about candidates’ was applied where an examiner inferred student characteristics (e.g. ability, lazy, thoughtful) or inferred how a student has approached the task from the student’s response. For example: *‘she is someone who knows she has got a target, the question, but I think she is doing that to make up for her lack of any knowledge’*. Often, assumptions about candidates were 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. Sometimes verbalisations suggested that the examiner was trying to understand why the student has produced the kind of response that they have (e.g. *‘so basically what she has done is run out of time in her exam and sadly right at the end she has come up with some really good ideas, but she hasn’t had time to develop them which is a pity’*). Additionally, a few instances of this code were about experiences that the examiner assumes the student has had (e.g. that the student has been to an out-of-town shopping centre or has seen a particular video). 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:

- *‘it’s not a feeling that this person is a geographer’;*
- *‘this lad knows a lot, likes to write a lot’;*
- *‘my mind was thinking this is a bright boy but very lazy who doesn’t really know a great deal of information’;*

- *‘again this candidate is actually, I’d love to talk to them they seem to be rather sort of messy but with quite a good brain’*.

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 and referred to the student being *‘a bit more on the ball here, that’s good’*, writing *‘all he could think of’*, or providing *‘a nice unusual twist’* in his/her writing.

Whilst making assumptions about candidate characteristics would seem to be a potentially dangerous aspect of examiner marking behaviours, closer examination of such instances has revealed that as far as we can tell these reactions only sometimes feed into overall evaluations and are not inappropriate. Such behaviours seem to be part of an examiner’s way of pulling together their view of the student’s abilities so far and understanding why a student has written a certain kind of answer.

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 (e.g. *‘I think had she been able to complete she would have scored quite a high mark’*). 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 submission (a)). 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 as the following examples illustrate:

- *‘This is not going to be a better paper is it’;*
- *‘I think we are just going to get all they know about Hurricane (Mitch)’;*

- *‘one wonders whether this knowledge is going to be used to answer the question later on’.*

Predicting performance too early on in a response or script and then being compelled to stick to this view regardless of further evidence would be a dangerous marking strategy, hence its consideration here.

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. However, concurrent evaluations did not appear to influence later discussion of a response with the examiners’ views being malleable (as illustrated in this example extract: *‘his introductory paragraph does not fill me with confidence… but let’s move on and see’*) and strongly linked to the response content and how this relates to the marking criteria. Further to this, 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.

Again it seems that predicting further performance based on features observed so far in a script is not problematic because such behaviours are another part of an assessor forming an ‘image’ of the work so far. Expectations of further performance do not seem to lessen the care with which the rest of the response is considered or to bias judgements regarding appropriate marks.

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.

Reviewing the instances where examiners referred to aspects of the teaching received by candidates revealed that comments about teaching never led to concurrent or overall evaluations during either marking or grading. Possible characteristics of teaching seemed to be used by examiners to make sense of why certain aspects of responses occurred.

Examiners sometimes entered into a constructed dialogue with the student via the text indicating a degree of social engagement. Examiners appeared to talk to or about the student (rather than the student's work) 1.08 times per script on average during marking. This seems to occur to help make sense of student responses and meanings and is not related to evaluations in their own right. Whilst verbalisations involving talking to or about the student were sometimes related to an evaluation of an aspect of the student's work (e.g. *'well why don't you name one'*, *'so are you saying that some avalanches are caused without human interference?'*) such verbalisations were often part of reading, making sense of responses and scrutinising for meaning and quality and were not related to evaluation in their own right. The nature of extracts showing talk to or about the student was similar in grading though less frequently occurring.

Personal and affective reactions

Examiners sometimes showed affective (i.e. emotional) or personal reactions to features of students' work (Crisp, 2007; Crisp, in submission (a)). As with the social perception codes, the behaviours coded in this category do not represent the actual features being paid attention to by examiners, but examiners' reactions to features of student work. During marking, positive affect or sympathy towards the candidate (e.g. *'so good he is on target now, I'm really pleased', 'hasn't developed the idea which is a pity'*) 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. Where positive affect was associated with overall evaluations it could appear that an examiner’s emotional reaction is guiding the marking decisions rather than evidence of the marking criteria being met (e.g. *‘I quite like that so I will give that top marks’)*. However, looking at the verbalisations in context indicates that in fact the characteristics that lead an examiner to say that they ‘like’ the response are usually connected to geographical knowledge or skills or the student’s efforts to address the question asked (e.g. *‘but I like it there’s a nice feeling that someone is actually questioning the title’*). As such, positive emotional reactions to students’ work do not seem to be problematic.

In grading, evidence of positive affect was fairly infrequent (0.19 times per script on average or a total of 17 instances). 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. *‘this one as soon as I look at it I am not very happy’*, *‘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 (e.g. *‘so actually begrudgingly, I’d really like more detail, they are going to get into level 3 aren’t they’*, ‘*oh [sigh], I want to give it nought but there is something there isn’t there”*). Negative affective reactions were infrequent in grading (0.10 times per script on average, or a total of 9 instances). 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. Most cases were reactions to aspects of the student’s writing such as an incorrect recall of something (e.g. one student wrote about a ‘tortoise butterfly’) or an amusing statement (e.g. *‘yeah retired people don’t want to work [laughter]’*). Only 6 instances were linked to concurrent evaluations and none to overall evaluations. The concurrent evaluations tended to occur where the student gives certain kinds of factually incorrect information perhaps about specific places which are then evaluated as incorrect (e.g. *‘Bournemouth is in Dorset it’s not in Hampshire [laughter]’*). Amusement and laughter were infrequent in grading (0.09 times per script on average or 8 instances) 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 or something was missing from their response (e.g. not including examples or not being very clear) or their response was not appropriately targeted to the question. In grading frustration was infrequent (0.01 times per script on average or a total of 10 instances). 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 ability to achieve the task as apparent in the student's work 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. This was particularly frequent in ICT (19.5 times per coursework piece on average) where the teacher's evaluations seemed to be very closely driven by the features of student work required by the mark scheme. In English the marking guidance was referred to 3.5 times per coursework folder on average perhaps reflecting the more holistic banded nature of the marking scheme.

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).

Although this analysis is still in progress, 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

There are of course some limitations to this study in terms of the use of one main subject area in one type of qualification and the fairly small number of assessors involved. However, the pilot data from GCSE coursework marking give some indication that the findings are likely to generalise beyond A level geography, although this is only based on two assessors. Generalisation to different subjects, qualifications, types of assessments and assessment systems cannot be assumed without further research.

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 & Simon, 1993) is problematic, particularly as some types of cognitive processes occur below consciousness. For example, dual processing theories (Sloman, 2002; Kahneman & Frederick, 2002) propose that some reasoning processes are automatic and associative and occur below consciousness. Therefore, we can not 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 focused 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. These included the quality of students’ language use (sometimes this was not an explicit criteria for marking) and evaluations of response length. There were also indications of assessors having affective and personal reactions to student work (e.g. like or dislike, amusement or frustration) or having social perceptions of students (e.g. predicting further performance, making assumptions about candidates). These reactions will have been a result of certain features of student work but it would not be appropriate for such reactions to influence assessment judgements.

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, 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.

In the case of reactions to student work that might not be considered appropriate influences on judgements (i.e. social perceptions and affective reactions, e.g. *'I quite like that so I will give that top marks'*) close analysis indicated that most instances were actually caused by features of the student work that were intended to be evaluated. Assessors again seemed to be in control of their social and affective reactions such that only features intended to be considered in marking were used.

Considering the findings with respect to Sadler's (1989) notion of manifest and latent criteria, we might think of criteria set out in the mark scheme as normally those that are 'manifest' during at least some of the marking process. Criteria such as language use (where not an intended criterion) and response length might be part of the latent criteria that sometimes come into use (though in a controlled way). It seems that the 'pool' of latent criteria available and sometimes brought into use does not extend far beyond the range of criteria set out in the marking guidance.

This may suggest that an analytic type of judgement model may be more appropriate than a configurational model given that the criteria used for assessment do not seem to be drawn from a wide pool and that specific features are noted along the route to an overall judgement (rather than afterwards). However, a more detailed analysis of all legitimate features of work being attended to and used in assessment would be needed in order to determine this. It is also likely that even if an analytic model was found to underpin the judgements, that it would be via a fairly flexible rather than rigid procedural use of various criteria given that examiners were found to vary in the behaviours apparent in their marking (see Crisp, 2007; Crisp, in submission (a)).

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. Further research and analysis would be needed to confirm this. This would 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 partly ambiguous response, or variations in the interpretation of aspects of the mark scheme. Again these issues would require further investigation to ascertain their contribution.

This research has shown that verbal protocol analysis can provide a method to evaluate the quality of assessments made by judges by investigating whether the processes reflect the intended criteria of evaluation. This is in line with the findings of Sanderson (2001). It may be possible to remedy any discrepancies uncovered using training. Verbal protocol analysis can also allow us to investigate issues such as the processes associated with inaccuracy or severity and leniency and to compare the processes involved in marking between different types of tasks (e.g. short questions versus essays or coursework, see Crisp, 2007), between different individual assessors or types of assessors (e.g. teachers versus examiners), or between different aspects of the assessment process (e.g. marking scripts versus deciding on grade boundaries).

While the limitations of the methodology used should be remembered, this research has given us insight into whether assessors focus on appropriate criteria during assessment judgements. This is important as it could impact on the quality and reliability of assessments. The data are consistent with the view that the judgement processes involved in the assessment contexts investigated closely rely 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.

References

Berger, P., & Luckman, T. (1967) *The social construction of reality: a treatise in the sociology of knowledge* (London, Penguin).

Boaler, J. (1999) Participation, knowledge and beliefs: a community perspective on mathematics learning, *Educational Studies in Mathematics, 40*, 259-281.

Broadfoot, P. (1996) *Education, assessment and society: a sociological analysis* (Buckingham, Open University Press).

Cresswell, M. J. (1997) Examining judgements: theory and practice of awarding public examination grades. 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 submission (a)) Exploring the nature of examiner thinking during the process of examination marking, *Cambridge Journal of Education*.

Crisp, V. (to be submitted) A tentative model of the judgement processes involved in examination marking.

Delaney, C. M. (2005) The evaluation of university students' written work, paper presented at the British Educational Research Association Annual Conference, Glamorgan, available at: http://www.leeds.ac.uk/educol/documents/149754.doc (accessed 9 January 2007).

Dreyfus, H. L., & Dreyfus, S. E. (2005) Expertise in real world contexts, *Organization Studies, 26*, 779-792.

Einhorn, H. J. (2000) Expert judgement: some necessary conditions and an example, in: T. Connelly, H.R. Arkes, & K.R. Hammond (Eds) *Judgement and decision making: an interdisciplinary reader* (2nd ed., pp. 324-335) (Cambridge, Cambridge University Press).

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).

Filer, A. (2000) *Assessment: social practice and social product* (London, Routledge and Falmer).

French, S., Allat, P., Slater, J., Vassiloglou, M., & Willmutt, A. (1992)

Implementation of a decision analytic aid to support examiners' judgements in aggregating pairs of components, *Journal of Mathematical and Statistical Psychology, 45*, 75-91.

Gernsbacher, M. A. (1990) *Language comprehension as structure building* (Hillsdale, NJ, Lawrence Erlbaum Associates).

Gilovich, T., Griffin, D., & Kahneman, D. (2002) *Heuristics and biases: the psychology of intuitive judgments* (Cambridge, Cambridge University Press).

Gipps, C. (1999) Socio-cultural aspects of assessment, *Review of Research in Education, 24*, 355-392.

Greatorex, J., & Suto, W. M. I. (2006) An empirical exploration of human judgement in the marking of school examinations, paper presented at the International Association for Educational Assessment Conference, Singapore, available at: http://www.iaea2006.seab.gov.sg/conference/download/papers/An%20empirical%20exploration%20of%20human%20judgement%20in%20the%20marking%20of%20school%20examinations.pdf (accessed 9 January 2007).

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, Hampton 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).

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

Massey, A., & Dexter, T. (2002) An evaluation of Spelling, Punctuation and Grammar assessment in GCSE, paper presented at the British Educational Research Association Annual Conference, Exeter University.

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*.

Nisbett, R. E., & Wilson, T. D. (1977) Telling more than we can know: verbal reports on mental processes, *Psychological Review, 84*, 231-259.

Qualifications and Curriculum Authority. (2007) *Code of Practice* (London, QCA).

Sadler, D. R. (1989) Formative assessment and the design of instructional systems, *Instructional Science, 18*, 119-144.

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.

Shay, S. (2004) The assessment of complex performances: a socially-situated interpretive act, *Harvard Educational Review, 74*(3), 307-329.

Shay, S. (2005) The assessment of complex tasks: a double reading, *Studies in Higher Education, 30*(6), 663-679.

Sloman, S. A. (2002) Two systems of reasoning, in: T. Gilovich, D. Griffin, & D. Kahneman (Eds) *Heuristics and biases: the psychology of intuitive judgment* (Cambridge, Cambridge University Press).

van Dijk, T. A., & Kintsch, W. (1983) *Strategies of discourse comprehension* (New York, Academic Press).

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).

Wenger, E. (1998) *Communities of practice, learning meaning and identity* (Cambridge, Cambridge University Press).

Working Group on 14-19 Reform. (2004) *14-19 Curriculum and Qualifications Reform: Final report of the Working Group on 14-19 Reform (Tomlinson Report)* (Annesley, DfES Publications).