How Do Assessors Mark? The Process of Assessing Written Work Produced by Students in Higher Education

Submitted by Calum Milne Delaney to the University of Exeter
as a thesis for the degree of *Doctorate of Education in Education*
June 2012

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(Signature).....

Abstract

Much research into assessment has concentrated on its role in learning and educational practice, issues relating to objectivity and reliability in assessment, and the political and policy implications of assessment more generally. The means by which assessors arrive at their judgement has received comparatively little attention and remains obscure. There has been a focus on factors relating to the product rather than the subjectively experienced process of assessment. A greater understanding of the process is important for the validity of assessment and its wider consequences for students and others.

The aim of this study was to examine how assessors conceptualise and carry out the assessment of discursive writing produced by students in a higher education context. Semi-structured interviews were conducted with experienced lecturers in health care subjects. The interviews and the data analysis were approached from within a hermeneutic phenomenological tradition, involving both description and interpretation. The participants' descriptions provided an analogue of what they thought they did cognitively as they assessed. These texts were then subjected to interpretation negotiated with participants to develop an understanding of the assessment process.

There were two main findings relating to how participants carried out the process of assessment. Firstly, they made use of a framework of meanings that appeared in part to arise from the practice of evaluating in terms of grade-bands. These were viewed as having categorical identities with discontinuities between them, as opposed to representing ranges within a continuous scale. The data suggested that there were changes in the aspects of writing to which assessors paid attention (content versus argument/integration and components versus the whole), and the kinds of judgements they made (quantitative versus qualitative), at different points along the grade band scale.

Secondly, the participants made use of six categories of processes during the course of performing an assessment. Some were objective and analytical while others were more subjective and integrative. They were not carried out sequentially, but appeared to be determined by the demands of the assessment task and to serve a function of simplification. The variety of processes within each category, their co-occurring usage and interdependence, and the selective use (or awareness) of processes by different assessors may help to explain some of the apparent complexity inherent in the assessment task, and the difficulty that experienced assessors demonstrate when trying to explain what it is they do and how they do it.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I wish to thank my supervisors for their guidance and encouragement, my participants for their interest and willingness to participate, and my wife for her unfailing support.

Examining: Reconciling Marks (2)

'Oh yes. Pretty straightforward this one, I thought. One of those which more or less marks itself. Definitely second class in my opinion. No hint of anything first rate'.
'Absolutely. Second class written all over it'.
'And, on the whole, I thought it was more or less in the *lower* second bracket. Rather lower-secondish feel to most of it'.
'No doubt about it. Very lowerish'.
'And really when you got down to it, rather middle lower-secondish. Not any higher than that – not as it were moving at all towards a borderline upper, but then on the other hand quite clearly a good distance away from anything resembling a borderline third'.
'Middle lower second?'
'Yes, I thought so?'
'Smack in the middle?'
'Yes. On the whole. Middle middle lower second'.
'Mmmm'
'You've gone a bit higher, have you? Sort of upper middle lower second?'
'No. Quite the contrary. I mean, I'm very much with you on the lower second and agree that it's not at all borderline in any way – but quite honestly I've gone much more towards the lower part of the middle – sort of lower middle lower second'.
'Yeees. I just don't think it's got the sort of quality that you'd expect from a good solid middle lower second'.
'I must disagree with you quite profoundly. I thought it had *just* that sort of quality. You know, a certain honesty which was very very middle lower. Not much depth of course – I mean you don't expect it at this level – but a certain weight. A certain weight'.
'Well, you surprise me. I found that it was just . . . well . . . may I speak frankly?'
'Please do. Please do'.
'Well, I found it just a little on the thin side of things'.
'Thin?'
'No, not "*thin"*. Thinn*ish.* On the thin side of things'.
'How very odd. Still, what's your actual mark?'
'I've gone straight for a 53 here'.
'Fifty-*three*?'
'Yes'.
'Plumb in the middle of the lower middle lower second? Not at all borderline?'
'No, I'm afraid not. And you?'
'Well, as I say. I'm right in the middle middle here. Smack on 55'.
'That looks like a straight disagreement then?'
'Yes. I'm afraid so. No two ways about it'.
'One for the external?'
'Yes, I think so'.

(Laurie Taylor, 1986, *Professor Lapping Sends His Apologies: The Best of Laurie Taylor,* Stoke-on-Trent: Trentham Books)

Chapter 1: Introduction

Motivation for the research

My motivation for carrying out this research stems from three sources. The first is my own puzzlement at what it is I actually do when I assess written work. The second is the repeated assertion in the literature that the assessment process is obscure, poorly understood and that a better understanding is needed of the judgement involved. The third is the repeated finding that the judgements made of a piece of writing can vary across assessors and across repeated assessments of the same piece of work by the same assessor. In this section I will provide an account of the personal and professional context for this research that locates it within my practice as an educator, and an explanation of the rationale for the research within the broader context of assessment.

I am conscious that I am doing several things when I am reading, interpreting, placing a value on and assigning a mark to a piece of written work, but that these are ill-defined, appear to merge into one other, and some are more obvious for some pieces of work than others. The focus of my assessment can also change from essay to essay within a batch of scripts. Thus there are a number of variables that I appear to be juggling and the relative importance of these variables is constantly altering. I also need to weigh up these variables in some way to arrive at a final mark or grade for a piece of writing. The majority of my teaching has been into the undergraduate pre-registration education of speech and language therapists and audiologists, and although I have been assessing the written work of students for almost 30 years I am still uncertain of what it is I do or what it is I am supposed to do.

Early on in my career I approached colleagues within my department and the wider university to ask them about the assessment process and how we were meant to go about it. My focus initially was on the reasoning behind the grading scale. In South Africa, where I began my university teaching career, this consisted of a pass mark at 50% followed by intervals of 10%, 10%, 5%, and 25% designated as a third, lower

second, upper second, and first respectively – not too dissimilar to the more symmetrical third, 2:2, 2:1 and first (starting at 40% and having intervals of 10%, 10%, 10% and 30% respectively) found in most universities in the United Kingdom. As a student who had moved through the same system, I had assumed that there was some reason or logic to it. No one I spoke to knew of any logic and suggested that it was "just the way it was". With respect to the process of forming a judgement and assigning a mark to a piece of work the responses were similarly vague. My experience was echoed 30 years later by the comment of one of my participants who said "there is no system telling you how to do this" and "when you ask for the book of instructions on how to do it [the establishment] says 'well you're the professional'". More senior colleagues helped me though, by explaining why they might award a particular mark, or a mark higher or lower than mine. I joined in the verbal fencing described by Shay (2004) as I negotiated my interpretations with my colleagues while trying to avoid threatening each other's integrity. This increased my awareness of how academic judgement can be socially constituted (Bourdieu, 1988). Although I continued to carry out my assessment practices with less concern (I must have been getting it approximately right as I was never seriously challenged about it), I nevertheless continued to think about it. I also had a persisting impression that there might be a certain amount of arbitrariness to the process, or that the complexity of what actually goes on is poorly represented by what can be observed on the surface.

Previous research on assessment has recognised both the high stakes implications of summative assessment and how formative assessment is interwoven with the learning process (Wolcott, 1998; Haines, 2004). Written work is used as the basis for assessment in a wide variety of educational contexts. The assessor's view of a piece of writing thus plays an important part in both the classification or certification of the student as well as the ongoing process of interaction with the student that supports his or her learning. While the approach to assessing written work in schools has focussed on increasing standardisation to promote reliability (Laming, 2004; Crisp and Johnson, 2007), at the university level a more informal and local approach is prevalent (Shay, 2005; Jawitz, 2009). This may be related to fewer constraints being placed on work considered acceptable in a university context (Bloxham, Boyd and Orr, 2011). Although there is a

recognition of a need for a better understanding of assessment in education, several researchers have suggested that there are aspects of the process, involving the assessor's judgement, that remain obscure (Huot, 1990, 2002; DeRemer, 1998; Lumley, 2002) and opaque (Wyatt-Smith and Castleton, 2005). This is in spite of attempts to objectify and standardise assessment by using analytic rating scales and rubrics. Huot (1996, 2002), in his discussion of assessment validity, suggested that the nature of the judgement involved in writing assessment needed to be incorporated into a theoretical understanding of the process. Although the assessor's task appears to be one of comparing aspects of a piece of written work to criteria listed in guidelines or rubrics (Broad, 2000), and/or determining where these aspects might fall on a rating scale, previous research has suggested that the process of doing so is complex (DeRemer, 1998) and that the relationship between the writing and its judgement is unclear (Lumley, 2002; Crisp, 2008a).

In addition to this, and in spite of attempts to improve reliability in assessment, some studies have reported poor consistency, or variations in consistency, of marking or grading across different assessors and assessments (Baume and Yorke, 2002; Baird, Greatorex and Bell, 2004; Shay, 2005; Read, Francis and Robson, 2005; Bloxham, 2009; Hunter and Docherty, 2011). One reason offered for this has been the complexity of the process arising out of the number of variables involved. Assessors also manage these variables in different and sometimes idiosyncratic ways which reflect how they conceptualise and make sense of the process of reading, evaluating and judging writing, and how they locate this activity in the broader context of education and their practice within it. Simplification or reduction of these variables, in the interest of increasing consistency, changes the nature of the process. Conceptualising research into assessment in similar ways decreases the likelihood of it producing results that accurately reflect the process or explain the inconsistencies (Broad, 2000; Johnston, 2004). The aspects of writing assessment that are poorly understood and that might help to explain the variable and apparently complex ways in which assessors perform the task, are those that are less accessible to the researcher and difficult to identify and describe. The nature of the data therefore presents methodological difficulties and the integrity of the results will be vulnerable to the greater degree of interpretation needed

to generate them. However, acceptance of these difficulties and limitations is necessary if these lesser examined aspects of assessment are to be explored.

The broader context of assessment

There are two factors driving the increase in the visibility and importance of assessment in educational discourse. These are the increasing centralisation and domination of educational provision by governments and governmental agencies (e.g. Broadfoot, 1996; Isaacs, 2010), and the concomitant and related casting of education in economic and commercial terms (Naidoo and Jamieson, 2005; Haggis, 2006; Walsh, 2006) where the results of assessment serve as both a product as well as a means of evaluating the worth of the educational provision. The 1988 Education Reform Act and the introduction of a national curriculum consolidated centralised governmental regulation of education in schools in England, Wales and Northern Ireland. The linking of curriculum and assessment became more overt with the forming of the School Curriculum and Assessment Authority in 1993 by the merging of the National Curriculum Council and the School Examinations and Assessment Council. The thinking behind a national curriculum and its attendant assessment regime was that it would promote a raising of educational standards, which reflected an increasing prominence being given to the role of education in the global competitiveness of a state (Walsh, 2006). The linking of curriculum and assessment also reflected a desire to specify and measure the outcomes of education and to set targets for education that could be related to those measurements (Broadfoot, 1996). As a consequence of this the results of assessment became increasingly public. In addition to providing a measure of individual progress and the overall standard of education, it also began to serve as a means for measuring the performance of teachers and schools and for enforcing accountability (Marshall, 2007; Stobart, 2008). This has had a cumulative effect on teaching provision - teaching to the test and an increasingly rigid approach to the curriculum (Isaacs, 2010; Walsh, 2006). An additional feature of the current emphasis on accountability has been the increasing control by government of the products of education through its various agencies. This has given the government a larger stake in

the outcomes of education, as it too has come to be evaluated in terms of the measured outcomes of its policies (Isaacs, 2010).

The influence of government policy on assessment in higher education has been less extensive in terms of direct intervention, but the broader context and in particular the marketisation of higher education has resulted in an increased scrutiny of assessment practices. Although less overt, many of the arguments for the audit and accountability functions of assessment in primary and secondary education are found in higher education and have begun to exert a similar influence on the discourse around standards (and raising standards) and value for money (Boud, 2007). Since the introduction of top-up fees in 2006, this discourse has involved the participants in education as well as the state. Together with the continuing selection and certification functions of assessment, and the persisting assumptions about its diagnostic and predictive capabilities (Stobart, 2008), these have contributed to the high stakes of assessment for higher educational practice. A consequence of this has been a focus on the outcome of assessment rather than the process of learning, thus motivating student behaviour that fits that focus (Boud, 2007). The importance accorded to certification and audit has also diverted attention away from its potential developmental function (Price, O'Donovan, Rust and Carroll, 2008).

While there has been an increasing focus on the outcomes of assessment in society at large, there has also been a shift in the practices of assessment in the education community. Ecclestone (2002, 2003) has described how norm-referenced criteria have been giving way to criterion-referenced criteria over the past 30 years or so, although this only began to influence practices in higher education in the 1990s (Ecclestone, 2001). In part this may be seen as a reflection of the changing view of the role and purpose of education in society and the economy, and its link to work and global competitiveness (Naidoo and Jamieson, 2005). It aligns with an emphasis on outcomes and the measurement necessary to provide evidence of those outcomes. There has also been a discourse around the importance of de-emphasising summative assessment while placing a greater emphasis on formative assessment in order to better facilitate learning (e.g. Knight, 2006; Boud and Falchikov, 2007). Price et al. (2008) provide a sample of

the views expressed in this debate. The purposes currently served by assessment require an emphasis on measurement and quantification that continues to undermine its formative potential, highlighting how the purposes served by assessment determine what is learned and how it is learned. This may provide an insight into the difficulties encountered in implementing formative assessment and encouraging meaningful learning that are repeatedly described in the literature (e.g. Knight, 2002; Rust, 2007).

On account of the focus on the measurement and audit roles of assessment there has been an understandable emphasis on the reliability of that measurement. The effect of this has been to proceduralise assessment, and Crook, Gross and Dymott (2006) have suggested that this can mask what may actually be going on. This may relate to both the declared educational purposes of the assessment (what it is that the students do) as well as to what it is that assessors do as part of the process. Citing Brown and Duguid (2000), Crook et al. also suggested that one effect of this is to conceptualise assessment as a process rather than a practice (as described by Wenger, 1998, for example). A consequence of this is for the enterprise of assessment and its research to focus on the procedural, diverting attention away from the sorts of practices that may be occurring and that may contribute to a more meaningful understanding of the totality of the activity. Thus in addition to assessment being appropriated for purposes of certification and accountability, a focus on process may be limiting the way in which research into assessment is being conceptualised.

As a result of the place that assessment holds in public discourse, much of the research on assessment has focussed on its public face and the aspects of it that drive public interest. These aspects relate to standards and standardisation, the specification of criteria for assessment and how these should be applied, and procedures for normalising the application of criteria across different assessors and occurrences of assessment (Shay, 2008; Price, 2005). They also reflect the belief that the assessment process should be capable of greater objectivity. However as Wyatt-Smith and Castleton (2005) pointed out, this public focus has been at the expense of examining the private process of judgement that is experienced by assessors. They and others (e.g. Price, 2005; Crisp, 2008a; Shay, 2008; Bloxham et al. 2011) note that the ways in which assessors arrive at

their judgements within frameworks of technical procedures and criteria have been little researched. The private aspects of assessment are often placed outside the scope of assessment research. This is partly a consequence of a prevailing ideological orientation that does not easily accommodate subjectivity, but it is also because the private aspects of assessment are not as easy to examine as the public aspects. This raises methodological issues for such research and these will be discussed further in Chapter 3.

The aim of the present study is to contribute to an understanding of the assessment process specifically in relation to undergraduate student writing. The aim is also to examine in particular the aspects of the process that are private or internal to the assessor, and to endeavour to recognise some of the methodological difficulties that have been less frequently tackled in previous research. It is anticipated that by doing so it may be possible to bring additional data and insights to bear that might contribute to a more integrated explanation of the assessment process. The research therefore has the potential to contribute new data and interpretations of that data, to address some of the methodological and evidence gaps in previous research, and to offer an explanation of a specific aspect of assessment practice. These may contribute to a more complete account of the inconsistencies identified in that practice.

In Chapter 2 a review of relevant literature examines previous studies on assessment in terms of both how they contribute to an understanding of assessment and some of the methodological issues they raise. Chapter 3 provides a description of my research perspective and the methods employed in the study. Chapter 4 describes the findings and how these are supported by the data. Chapter 5 provides a discussion of the findings and how these relate to previous work. In my conclusions in Chapter 6, I provide an overview of the implications of the findings and a discussion of their contribution to an understanding of the assessment of discursive written work.

Chapter 2: Literature Review

Introduction

Wyatt-Smith and Castleton (2005) have noted that judgement in the assessment of discursive written work not only relates to assigning a mark or grade to a piece of writing, but may also involve decisions about criteria, about the extent to which aspects of the work satisfy the criteria, and about how criteria should relate to points along the grading scale. These decisions relating to judgement and criteria reflect deeper assessment-related variables that are often excluded from consideration in studies of assessment. Alternatively they are described as involving professional judgement or connoisseurship and are explored no further. In addition to aspects of assessment that concern judgement and the criteria that support judgement, several researchers have focussed attention on a feature of assessment that involves the reading of assessment pieces and arriving at a sense of the meaning of the writing. As discussed by Read, Francis and Robson (2004), the meaning of a piece of writing is not immutable and is constructed by the reader. Mullins and Kiley (2002) provided a description of their participants' approaches to their reading and the sorts of questions they had in mind to guide their sense-making, and Crisp (2010b) incorporated a reading phase into her model of assessment judgement. When considering the more private or internal aspects of assessment therefore, there are three potential areas of interest that might be investigated. The first involves the assessor's approach to the assessment task in terms of how they go about organising the process and how they read and make sense of the written texts. The second relates to aspects of the criteria and rubrics against which the writing is compared, and the assessors' perceptions or cognitive constructions of these. The third area relates to how assessors form an opinion of the value of the written work and make a judgement of its value in order to arrive at the mark or grade. In reviewing literature relevant to the assessment of written work I will organise the discussion around these three themes: meaning-making, criteria and judgement.

Meaning-making

In examining the behaviour of assessors, several studies have suggested that a distinction might be made between the assessor's understanding of the meaning of the writing and the judgement they subsequently make of it in order to award a mark or grade (e.g. Lumley, 2002; Crisp, 2008a). The judgement concerns the appropriateness of the meaning in relation to the topic or assessment task, and the effectiveness with which the meaning has been communicated. Prior to making a judgement it is necessary for the assessor to arrive at a sense of the meaning of the writing. Both the processes of arriving at the meaning, and a judgement, of a piece of writing are often referred to as evaluation. A consequence of this is to make the possible distinction between the two less evident, or to suggest that that they are not distinct and are interwoven in a single more complex process. The model described by Milanovic, Saville and Shuhong (1996) provided an example of this. It incorporated reading (to establish the overall level of comprehensibility) as part of the decision-making process in arriving at the final mark in the assessment of a piece of writing. In this study the authors discussed their interpretation of their data in terms of aspects of judgement rather than meaning-making. The assessment of relevance, coherence and the development of the topic were subsumed as part of the process of rating, but these might equally be viewed as contributing to the assessor's understanding of the meaning of the writing prior to a judgement being formed about it. The model thus de-emphasised meaning-making in assessment and suggested rather that it was part of the judgement process. Pollitt and Murray (1996), in a study that examined assessors evaluating students carrying out an oral task, observed that assessors tried to maintain a distinction between their qualitative appreciation of the performance and the quantitative evaluation they placed on it. An aim of their study was to try to develop an assessment scale that would combine the two aspects, thus again viewing them as part of the same process.

More recently however, researchers wanting to understand how assessors carry out assessment have recognised the possible importance of maintaining a distinction between meaning-making and judgement. In a study examining how assessors used

assessment rating scales, Lumley (2002) identified three broad types of assessment behaviours in his initial analysis. Two of these were reading behaviours and rating behaviours. (The third type concerned behaviours related to managing the assessment process). He likened these to the interpretation versus judgement strategies identified by Cumming, Kantor and Powers (2001, cited by Lumley, 2002). Mullins and Kiley (2002) studied the processes and judgements involved in the assessment of postgraduate research theses. Although much of their focus was directed towards the criteria and contextual features of assessment, they also explored how assessors worked through the text and formed their impressions of the writing. They identified a series of questions that their participants used to interrogate the writing. Some of these questions related to their conceptualisations of the topic, logical connections between parts of the thesis and explanation on the part of the writer, and the presence of an argument. Similarly, the impressions of the writing described by their participants related to whether they thought there was coherence in the writing, and whether it was written in a way that appropriately located it in its field. Crisp (2008a), in an analysis of 'think-aloud' commentaries produced by her participants while assessing A-level geography examination scripts, identified several reading behaviours. These included mechanical reading behaviours (e.g. reading aloud, summarising, paraphrasing). However she also identified behaviours characterised as scrutinising for meaning, scanning for something particular, making inferences, noting ambiguity and showing uncertainty with the writer's examples. She suggested that these behaviours might have represented an impaired understanding of the writer's meaning. An alternative interpretation might be that they were representative of processes employed by the assessor to make sense of the writing in order to ensure that the assessor did understand the writer's point.

In a small study, which examined participants' 'think-aloud' commentaries while marking undergraduate student essays, I also identified an apparent distinction between making sense of the writing and forming a judgement of its value (Delaney, 2005). With respect to the former, there appeared to be two aspects to how the assessors approached the writing. The first concerned the way in which they interacted with the writing. In addition to comments suggesting that an element of the writing was good or poor, the assessors offered a rationale for why they thought so. At times they also

disputed or argued or expressed exasperation with the points or arguments put forward in the writing. My interpretation of this behaviour was that it was as if the assessors were engaging in a conversation with the student through the writing. A similar observation of 'pseudo interpersonal relationships' was reported by Suto and Greatorex (2008b, p. 230). This interaction with the writing was less apparent in relation to the factual content of the essay, and more evident in relation to the thinking underlying that content. It also emerged when their comments suggested that they were trying to make sense of or interpret the writer's meaning. The second feature of how the assessors responded to the writing was that some aspects of the writing appeared to be easier to evaluate (in the sense of discerning the meaning) than others. With respect to the content aspects of the writing (the information or 'facts' being represented) the assessors appeared to be able to evaluate them more directly, and easily or quickly, on the basis of what was presented on the surface of the writing. However, they appeared to find it more difficult to evaluate the writer's thinking about the content elements as the thinking was not immediately apparent from the surface presentation of the content. The assessor needed to interact with the writing, and interpret its meanings, in order to try to arrive at an understanding of the writer's thinking about the content.

Lumley (2002) also identified meaning-making on the part of the assessor in the responses of his participants when deciding how to score writing using a holistic scoring rubric. They expressed "expert reactions, complex thoughts and conflicting feelings" (p. 267) in their 'think-aloud' commentaries while reading texts produced by English second language learners. He described his participants as forming complex impressions of the writing which they then needed to reconcile with the descriptors in the rubric. He suggested that this was likely to be more difficult when the writing was more problematic. His interpretation of this finding was that the application of the rubric was more difficult when the meaning of the writing was less certain (and by implication easier when the content more readily matched the rubric). He viewed this as a difficulty inherent in the use of rubrics for evaluating writing. However, an alternative approach to these kinds of data might be to view them as offering evidence for the existence of a process of meaning-making separate from that of judgement, and of content more easily discerned and matched to a rubric versus deeper less discernable

meanings that require more effort on the part of the assessor to make sense of them.

Whether or not meaning-making is part of the process of assessment, several authors have argued that the meaning of any piece of writing is not objectively represented by or contained within the writing. Rather it is constructed by the reader and open to a variety of interpretations (Read et al., 2004; Huot and Perry, 2009). Scott (2005) described reading as a process of interpretation or making sense of text and as something apart from assessment. She viewed openness to possible meanings as being an important characteristic of the reader. She also noted a distinction that might be drawn between words and the concepts they might signify. Individual words can have several meanings (particularly when placed in different contexts or when there are unstated assumptions or understandings) and some concepts can have more than one word to represent them. Thus there are surface representations and underlying meanings in any text, and the relationship between them is not always immutable - the relationship has to be established by the reader. Contreras-McGavin and Kezar (2007) also emphasised the interpretation task of the assessor. Rather than focussing on this activity as permitting the assessor to understand the meanings of the writer, they viewed it as the means by which the assessor discerned the learning and development of the writer. Importantly they saw it as a mechanism whereby the assessor tried to understand the student's understanding rather than the extent to which the writing did or did not fit the assessor's conceptualisation of what should be important. The writing was therefore a representation of the student's understanding of the assessment task. Scott, and Contreras-McGavin and Kezar, sought to emphasise the qualitative and interpretative nature of assessment, and conceptualised written work as a representation of the student's learning or understanding with its own internal logic and meaning.

Read et al. (2004, 2005) suggested that the meanings constructed of written work by assessors were influenced by factors relating to their social positioning, cultural identity and their views of the world. They emphasised the point made by Gadamer (1993) that the meaning of a text arises both from the text produced in the past, and the sense that is made of it in the present by the reader. Shay (2004, 2005) and Jawitz (2009) described, from the point of view of assessors in higher education, the complex social context

within which assessment takes place. An important aspect of this was the local effect of that context on the practices of their participants. Like Scott (2005), Knight (2006) additionally commented on the social context of the writer. He described assessment as 'doubly contexted' (p. 435), reflecting both the context within which the writer produced the writing and that in which the assessor produced his or her evaluation of the work. Baume and Yorke (2002), discussing assessment reliability, asserted the importance of shared understanding or inter-subjectivity between assessors and assessed in relation to the purpose of the process. These studies support the possibility that there are three potential sources of meaning, or contexts for meaning, for a piece of written work. These are the meaning intended by the writer, the meaning discerned by the reader and an aspect of meaning contained within the text itself. Each of these may be influenced by a variety of variables and each may give rise to a variety of interpretations of the text. It is thus not unreasonable to suggest that the meaning of a piece of written work may be contestable, and that it may be necessary for an assessor to go through a process of arriving at some sort of construction of at least one meaning before being in a position to form a judgement of the value of the writing. This places assessors in a hermeneutic role, and an understanding of their assessment practices will necessarily involve an examination of the personal and social factors that contribute to their interpretations of the writing they assess.

These interpretations then permit comparisons to be made between the writing and the criteria that organise and give meaning to the assessment, and judgements to be made about how the writing fits those criteria. Thus the process of meaning-making may be a precursor to the application of criteria and the forming of a judgement. It will therefore be important to examine the data in the proposed study for evidence of such meaning-making as a component or stage in the process of assessment.

Criteria

Introduction

The second area of interest relating to the private or internal aspects of assessment

concerns the criteria that assessors use as part of the process of judgement of written work and against which the writing is compared. The assessment process involves the assessor forming an impression of the writing as discussed above, and then his or her role appears to be to reconcile that impression with the criteria (Huot, 1990; Lumley, 2002; Calvert, 2005). On the face of it this aspect of assessment might seem uncontroversial, particularly in the context of the body of research directed towards criterion-referenced assessment (see for example Rezaei and Lovorn, 2010, for a summary). However this research has focussed on the aspects of criteria that can be described objectively. It has also sought to offer evidence to justify the appropriateness of such criteria in assessment. As was discussed above, the meanings of written work might be multiple, not immediately apparent from the surface presentation of the text and subject to a variety of interpretations. In a similar way, the criteria employed in assessment might be represented internally in the mind of the assessor in a variety of ways, and influenced by numerous variables (including those described above in relation to the meanings attributed to written work). There may also be other criteria employed by assessors that are not explicitly part of the assessment process and of which assessors themselves may not always be aware (Ecclestone, 2001). In the following discussion I examine some of the research and thinking relating to the criteria generally thought of in relation to assessment and more specifically to the private aspects of assessors' perceptions or cognitive constructions of criteria. I also examine how criteria might be used in assessment, and the meanings attaching to grades or marks and the scales within which these are located.

Assessment reliability

Broad (2000) and Huot (2002) provided summaries of the tensions between reliability and validity that have dominated theoretical discussion and practical approaches to writing assessment. This dispute has also been cast in terms of a conflict between public accountability and professional autonomy, between assessment as the application of explicit criteria within a defined protocol and something more complex and obscure akin to connoisseurship (Webster, Pepper and Jenkins, 2000; Knight, 2006). Much of the attention in writing assessment and research up until the 1990s focussed on its reliability and how this might be increased. In summarising some of this research,

Lumley (2002) noted that the primary ways of achieving reliability involved the refinement of scoring criteria specifications and the rigorous training of assessors to increase the level of agreement between them. Some attention was also given to the possibility that tighter specification of the writing task might also limit the variability inherent in its assessment (e.g Hamp-Lyons, 1991). Hoyt, Allred and Hunt (2010) described some of the procedures that are important to incorporate into assessment to improve reliability. In spite of these they nevertheless found room for further improvement. Their suggestions for this included further modifications to the criteria and the procedures they had employed. In a study that aimed to provide evidence for the effect on reliability of specific aspects of assessor training, Baird et al. (2004) found that the strategies did not result in an improvement in marking reliability.

In spite of attempts to improve reliability, some studies have reported that the consistency of marking across assessors and assessments is not very high. Wolf (1995) discussed some of these findings that related to vocational assessments, and Baird et al. (2004) provided a discussion of variables that adversely affect reliability, chiefly in the context of marking school-based assessments. Shay (2005) described her investigation into the reliability of the marking of undergraduate engineering final research projects, which were double marked. She reported correlation coefficients for the two percentage marks for each report of 0.71 and 0.72 for two successive years of reports. However when the two assessments for each report were compared in terms of the classification awarded (first, second and third), 62% and 51% of the reports for each of the years respectively had marks that differed by one or more classification bands. Similarly a study by Read et al. (2005) found that the grades awarded by 50 lecturers from 24 universities to two history essays varied from a 2:1 to a fail (approximately half of the assessors awarded both essays a 2:2). Baume and Yorke (2002), examining portfolio assessment in a higher education context, also found inconsistencies and that some aspects of the assessment showed greater consistency than others. Although their participants were relatively consistent in pass-fail discriminations, they showed less agreement on the quality of a passing or failing performance.

Broad (2000) discussed how the focus on reliability began to be challenged during the

1990s, when opponents argued that assessment agreement was forced or manipulated and necessitated a simplification of what was assessed. It also produced an unnatural way of reading that interfered with readers' engagement with the text, and the sense-making permitted to assessors ran the risk of reducing the authenticity and validity of that sense-making. In addition to this, Johnson (2004) commented on the implications of simplification for research. This conceptualisation of the assessment process might have restricted how it might have been understood, and limited the kinds of research that might have contributed to that understanding. Alternatively simplification might have obscured or altered the ways in which assessors arrived at their assessment decisions, contributing to limiting our understanding of the process (Huot, 2002). Broadening such research in ways suggested by Broad (2000) and Johnston (2004), and focussing more closely on the assessor, might offer a possibility for new directions for research and fresh insights into the process of assessment.

Sources of criteria

The explication of criteria to reduce personal or private aspects of the assessor's practice and increase objectivity has involved an emphasis on rubrics (primary trait, holistic, analytical or a combination of holistic and analytical), descriptors, scales and guidelines. These criteria relate to both the qualities of the written work and the system of grading, marking or scoring that is employed to quantify the worth of the work (Greatorex, 2001, 2002). Researchers have examined these criteria in order to identify aspects of these that might reduce their reliability. One finding has been that assessors sometimes interpret or use criteria in idiosyncratic ways, or they employ criteria additional to those that have been specified for the assessment (Webster et al., 2000; Ecclestone, 2001; Yorke, 2008). Hunter and Docherty (2011) described four categories of implicit beliefs or standards held by assessors in higher education. These related to what they expected in student writing (concepts like 'structure', 'argument' and 'clarity'), the meanings attached to these concepts, and the language used to express them, and more 'instinctive' (p. 113) beliefs relating to subtle aspects of writing such as tone and voice. Elbow (1997) also pointed out that assessors may not always be clear about the sources of their criteria.

Another feature of variations between assessors is that these are not always consistently present but may be triggered by features of the writing. Rezaei and Lovorn (2010) summarized some of the biases shown by assessors in response to aspects of the writing such as style, vocabulary, grammar and the presence of errors. They also described studies that showed that characteristics of the writer (e.g. sex and command of language, or general ability when the writer is known to the assessor) or of the assessor (harshness or leniency) could accentuate these biases. Ecclestone (2001, p. 308) described these aspects of assessor behaviour as 'unconscious interpretations' or 'compensations', where assessors incorporated contextual aspects of the assessment, the student or the student's performance into their assessments. The results of her study also suggested that this interpretive element of assessment was more prevalent in expert than in novice assessors. While the more textured interpretative assessment carried out by expert assessors might be seen as enhancing the validity of the process, the individual nature of the criteria and the extent to which they are incorporated will complicate the common understanding of the process across assessors, particularly where these include both novices and experts.

Writing as a generic skill

A considerable number of American studies on writing assessment relate to writing (or composition) as an independent endeavour or generic skill, expressed as competence or proficiency, largely divorced from the use of writing as a means of communication and expression of ideas. (See for example Beck and Jeffery, 2007, and Jeffery, 2009, for a summary of this perspective on the purpose of writing). Slomp (2012) suggested that this was related to ease of assessment. The criteria in these sorts of assessments emphasise procedural aspects of writing, and Erling and Richardson (2010) found that they were highly correlated with each other. They suggested that the criteria were all measuring the same underlying global construct, and that the assessors were unable to differentiate between characteristics of the writing measured by these criteria. It is also possible that it is difficult to assess in terms of one criterion without incorporating elements of others in the process. Scott (2008) offered a critique of standardised approaches to assessment where he argued that these produced a bureaucratisation of writing production and learning. Boud (2007) made similar observations. The focus on

the criteria results in the writing being shaped to suit the criteria. They become central to the teaching and learning of writing and begin to dominate the way in which it is conceptualised. The familiarisation of students with assessment criteria does not empower them with understanding, as suggested by Elwood and Klenowski (2002), but paradoxically makes them more subservient to the bureaucratic demands of the process because they have internalised this mechanism of control. Thus it is possible for criteria to dominate the writing enterprise in ways which undermine its purpose. This is likely to be particularly so where writing is viewed as a skill rather than a means of communication.

Writing as a means of communication

In contrast to a view of writing as a generic skill, investigations of the assessment of writing in universities place a greater emphasis on the purposes for such writing and the importance of local and contextual influences (Bloxham et al., 2011). This orientation reflects a more flexible positioning of assessors in relation to criteria, such that the criteria do not dominate the process in the way described by Scott (2008). It also makes the application of criteria less fixed and more contested (Condon, 2009). Condon described the purpose of writing in terms of academic competencies students needed to develop, and these involved understanding of factual content, presentation of content in order to convey the point of the writing, and the surface presentation of the writing. The criteria by which these were evaluated were: topic conception and clarity of ideas; focus and direction, argument logic, support and integration; and fluency, idiom and mechanics. Kreth, Crawford, Taylor and Brockman (2010) identified similar criteria or qualities of good writing, although these were viewed less as characteristics of the writing and more as representative of the thinking behind the writing. They also emphasised the communication function of the writing. In a study on the assessment of postgraduate thesis writing, Mullins and Kiley (2002) found that the criteria valued by their participants related to the intellectual depth, rigour and argument within the writing, the logical connection between the parts of the work, and the extent to which the writer had grounded it in the literature. A good thesis was one seen as showing scholarship (originality, coherence, autonomy and independence), a well-structured argument (which included conceptualisation, design, logic and structure), and evidence

of reflection and of having done what it set out to do. However, in spite of this expressed focus by assessors on the communication of thinking, logic and support for arguments, Kreth et al. (2010) noted that grammar, punctuation and spelling were the characteristics most frequently cited as identifying good writing. Bloxham et al. (2011) also reported a focus on 'surface' aspects of writing, but that their participants tried not to let these influence their decisions.

Criteria as properties of the written text

Reviewing research relating to the relationship between assessment criteria and learning, Elander, Harrington, Norton, Robinson and Reddy (2006) described writing as a means of demonstrating a result of learning, and assessment criteria as playing a role in the process of that learning through students' engagement with the criteria. Bloxham and Boyd (2007) characterised this as 'assessment as learning' (p. 15). Elander et al. (2006) described the criteria as relating to properties of the written text, and suggested that one purpose for assessment criteria was to enable students to develop their own tacit understanding of the academic writing enterprise. They summarised previous studies that had identified criteria for written work, ranging between 10 and 50 criteria. Reporting their own work, they had identified four core criteria for student writing. These were critical thinking, argument, use of language and structuring. They suggested that complex learning, being the construction of new knowledge and its transfer to practical situations (van Merrienboer, Kirschner and Kester, 2003), related most closely to the critical thinking and argument criteria. With respect to the language use and structuring criteria, these identified skills-based learning at lower levels of performance but identified evidence of deep learning in examples of better writing. Thus their work suggested that assessment criteria might relate to different levels of learning, or different types of academic functioning expressed through the writing (writing production skills versus more complex critical thinking, for example).

Elander et al. (2006) expanded on the meanings of these core criteria while discussing the learning they might assess. 'Critical thinking' included logic, clarity and accuracy of thought, scepticism towards ideas, identification of omissions in information and avoidance of bias and prejudice. They described 'argument' as the 'defining feature of

the essay' (p. 81) and suggested that the essay encapsulated the argument. They argued that varying (and often tacit) views about what might constitute a good argument might be a major contributor to the variability often encountered in writing assessment. 'Use of language' involved correctness (grammar, spelling and referencing), register (pragmatic appropriateness) and academic literacy (displaying the discourse of the academic community). In discussing 'structure', Elander et al. differentiated between form-driven structure (generic writing rules) and content-driven structure where the structure supported the logical presentation and organisation of the content. An important finding was that none of their criteria were representative of independent generic skills, and all required some integration of skills with more complex cognitive functioning. There was also a degree of conceptual overlap between the core criteria such that it was difficult to evaluate one without taking account of one or more of the others – they worked together. Kreth et al. (2010) made a similar observation. Two implications of this were that the abilities encapsulated by the criteria were inter-related and not distinct from each other, and that consequently an analytical application of the criteria during assessment was likely to be difficult. The extent of this inter-relatedness was partially influenced by the level of writing. Thus at a basic level of writing it was easier to identify separate aspects of criteria that were more skill-like, while as the writing became better its components became more integrated.

Writing as socialisation versus negotiation

Lea and Street (1998) provided a theoretical framework within which the research described above might be located. They identified three ways in which student writing might be conceptualised, based on how research has approached studies of writing. These are that writing may be described in terms of a set of skills, that it is a product of academic socialisation, or that it is a reflection of academic literacies. The first views writing as separate and easily identified technical skills that are generic across disciplines and contexts. They are often identifiable in the surface features of the writing such as grammar, spelling and perhaps broad aspects of the structure such as an introduction, a conclusion and the use of references. The academic socialisation perspective views writing within a broader social context as part of the enculturation of students into the practices of academic work, viewing the process from a constructivist

perspective. Writing is seen as relatively fixed and transparent, its meaning independent of the writer and the reader, and the academic culture is assumed to be relatively homogeneous. An academic literacies perspective views writing as involving a number of social and communicative practices in which different epistemologies and personal identities are negotiated. In addition to differences that may exist between students' and assessors' perspectives and institutional demands, writing practices need to vary to accommodate different genres, fields and disciplines. This perspective also acknowledges that the meanings made by students and assessors are contested and that their negotiation takes place in the context of a power imbalance. In delineating these three approaches or models of student writing, Lea and Street (1998) acknowledge that they need not be seen as independent of each other, and suggested that each subsequent approach emerged out of and in response to a recognition of limitations in the earlier approaches. It is thus not unreasonable to make use of all three when theorising and researching student writing and by extension its assessment.

The study reported by Lea and Street (1998) investigated views and practices relating to student writing and its assessment in two UK universities. They described a great deal of variation in writing requirements across disciplines, courses and assessors. They demonstrated that interpretations of the surface features of writing, as well as the meanings and understandings that underpinned them, could differ across disciplines, and that there were differences of opinion about what was considered legitimate. The assessor's view was also typically privileged over that of the student, as Beck (2006) has also pointed out. Lea and Street found that assessors rationalised their evaluations on the basis of the surface features of the writing rather than the context and the disciplinary expectations of the assessor. Assessors were able to identify good writing, but were less able to explain what defined good structure or argument or in what way a piece of writing lacked these characteristics. Elbow (1997) made a similar comment and suggested that perceptions of what constituted good writing might depend on the perspective of the assessor. These studies suggested that writing assessment could not be understood only in terms of generic characteristics but also was dependent on the discipline-specific understandings of the assessor which were not always made explicit to the student. From the point of view of the student the writing task was partly viewed

as requiring them to discern what was required of them, rather than just replicate a particular style of writing as an academic socialisation model might suggest. Slomp (2012) described this as metacognitive knowledge that students needed to demonstrate. Students also felt that guidelines did not deal with the most difficult issues, those around constructing writing that aligned with the expectations of the specific discipline, subject or assessor. Thus it may be important to recognise that writing in a higher education context consists of more than its surface features and the transparent objective transmission of information. It also involves potentially unstated meanings, understandings and expectations that contribute to the contested nature of assessment and to the complexity inherent in the process.

Departures from explicit assessment criteria

In addition to the types of criteria often used in writing assessment and the meanings accorded to these, there has been research into how assessors use or interact with the criteria. Webster et al. (2000) examined how formal assessment criteria were used to assess undergraduate dissertations in seven departments within a school of social sciences. Similarly to later findings by Lumley (2002) and Bloxham et al. (2011), Webster et al. noted that rather than use the criteria to arrive at a mark assessors used them to rationalise their marking decisions. Lumley commented on his participants' perceived obligation to couch their decisions in the wording of the criteria even though these were not used in formulating the decision. Mullins and Kiley (2002) and Bloxham et al. (2011) reported that a high proportion of their participants did not use the assessment criteria in a systematic way. Others checked the criteria, as advice or as a 'reality check', but mostly they depended on their individual experience and an independent sense of the standards in their discipline and the purpose of the writing task. An implication of this is that the perceived purpose of writing may constitute some of the criteria used in evaluation, or it may become blended with the criteria and potentially alter their meaning. The assessors' reliance on their individual opinions and individualistic approaches to conceptualising and using criteria also suggested that criteria might be employed in less than uniform ways.

Webster et al. (2000) described a number of ways in which their participants interacted

with the criteria. Examples of these were comments by the assessors that did not relate to the criteria, criteria that were ignored and others that were additional to the provided set, judgements influenced by wider value systems, differences (or differing proportions) in what assessors sought from the dissertations and differences in the relationship between assessors' comments and the marks they awarded. One possible reason for this may be related to how criteria are specified. Greatorex (2001) and Greatorex, Johnson and Frame (2001) noted that rubrics or criteria tend to be phrased in terms of positive characteristics or in terms of what is present rather than absent. This may be a consequence of a criterion-referenced orientation to assessment. As assessors may utilise a negative or absence dimension of a criterion in their assessment, the phrasing of criteria may colour or complicate their application, or partially explain the reported tendency of assessors to employ unstated or idiosyncratic criteria in their assessments. It also suggests that while such criteria bear some relationship to the learning outcomes an assessment is intended to assess, they may be less well aligned with what it is assessors do, or need to do, while carrying out the assessment process.

Webster et al. (2000) reported considerable variation in the meanings attached to words like 'analyse', 'concept', 'critical' and 'argument', and Woolf (2004) also commented on the language used to describe criteria. Their conclusions focussed on the negative implications of these findings for considerations of consistency and transparency, and consequently equity. While these are valid comments, these data might also be seen as providing some insights into how assessors actually view and use criteria in their assessment practice. With respect to the variable meanings assessors gave to different words, this may reflect the reality of a situation where words are nuanced or have multiple meanings and are potentially influenced by context. Rather than there being a one to one relationship between words and meaning, concepts may more usefully be thought of as existing as shifting nodes within matrices of defining words (Funnell, 2000). These data may therefore provide some indication of the potentially complex nature of the assessment process and suggest reasons why objectivity is difficult to achieve. They may also suggest why reducing these dimensions of assessment in the interests of promoting greater reliability may compromise its validity (Broad, 2000).

Recognising a tacit dimension to the assessment process, some researchers have sought to find ways of making this more explicit in the process of formulating criteria. Greatorex (2001, 2002) examined A-level assessors' perceptions of the writing they were assessing and the grades they used in their assessment. She did so using Repertory Grid Technique (Fransella and Bannister, 1977), with the aim of grounding descriptions of performance and grades within the meanings held of these by her participants. Two of her findings from these studies were that the characteristics of grades elicited were strongly associated with one another (Elander et al., 2006, echoed this finding), and that her participants tended to describe performances more broadly in terms of what writers had produced rather than in terms of specific characteristics. These data suggest that there may be difficulties inherent in making tacit knowledge explicit and/or the possibility that tacit knowledge involves more than can be captured by an assemblage of characteristics or criteria. The tendency of assessors to perceive characteristics as interlinked and to revert to broader conceptualisations of a performance may additionally suggest that attempting a narrow criterion focus in assessment may not fit well with how the evaluation process in writing assessment actually functions.

Variations in writing quality and criteria

Read et al. (2005) carried out a study that examined the perceptions of assessors marking undergraduate history essays in which they reported the positive and negative characteristics of the essays identified by their participants. The former included amount of reading, description and evaluation of subject matter, effectiveness of introduction, and effort. Characteristics identified as negative were presentation, grammar and spelling, and referencing. Characteristics of writing that were commented on both positively and negatively were understanding (or failure to answer the question) and relevance of information, argument, analysis and reflection and use of evidence, and structure and quality of writing. They suggested that assessors may identify certain criteria with poorer work and other criteria with stronger writing, while a third set of criteria might be used to evaluate writing across the range of quality. This supports the findings by Pollitt and Murray (1996) who suggested that assessors might use different criteria when evaluating performances at different ends of the grading scale. In my own previous study (Delaney, 2005) I also found that comments made by assessors when

evaluating better pieces of work were different to those made when assessing poorer writing. In the context of the findings by Elander et al. (2006) the negative criteria might be seen as relating to independent skills-based and generic aspects of the writing, while the positive criteria begin to incorporate the integration of more complex subject-based understanding but might still be considered relatively independently. The criteria relating to performance across the entire assessment range (with the possible exception of structure and quality) arguably involve the more complex and inter-related aspects of writing. This lends support to the possibility that criteria may be employed by assessors in different ways. There may be different types of criteria used for different aspects of the writing, and they might correspond to differences in the quality of the work. An examination of this possibility might contribute to a better understanding of the roles criteria play in assessment.

In a study that examined the descriptors used for the levels in the UK undergraduate and Master's degree structure, Greatorex (1999) suggested that the grades in a scale might be viewed in two ways. One is that they are hierarchical and represent a progression of steps, levels or stages of quality in relation to a particular characteristic, the characteristic being qualified as 'poor', 'fair', 'good' and so on for different grades. The second way is that some criteria for grades might be more salient than others at different points along the scale. Possible reasons for this are that salient criteria for poorer performances become less salient for better performances because they are less problematic for the student, or some criteria relate to characteristics that can only be found in better pieces of work. Greatorex described these two types of criteria as progressive and non-progressive respectively. The characteristics reported by Read et al. (2005) as positive or negative might be examples of non-progressive characteristics, whereas those that were commented on both positively and negatively might be viewed as progressive. Thus some criteria may exist on a continuum whereas others may uniquely identify a particular level. The latter type of criteria may complicate assessment in cases where writing incorporates non-progressive characteristics typical of more than one level in the scale. In cases like these, the criteria that take precedence might reveal something about the assessment process, or whether some criteria are viewed as essential while others are optional.

The grading system

Other researchers have examined more closely the grading systems used by assessors, and the interaction between the number of steps or levels in the grading scale and the number of criteria to be evaluated on the scale (Elbow, 1997; O'Donovan, Price and Rust, 2004). A typical grading scale used in most higher education contexts in the United Kingdom and in some other parts of the world consists of A, B, C and D pass grade-bands each divided into three sub-grades (designated A+, A, A-, B+, B, B- and so on). There is also an F or fail grade band, sometimes also divided into two or more sub-grades, resulting in a total of at least 13 levels between which an assessor is required to discriminate. Although rarely commented on, the meanings attached to the grade-bands awarded for individual pieces of work (A, B, C and D) are probably not unrelated to those attached to the classification system that qualifies undergraduate degrees in the UK (First, Upper Second, Lower Second and Third). One of the purposes of the grade and sub-grade scale is to permit relatively fine distinctions to be made between respective students' performances (Elbow, 1997). This can permit students to be ranked with some measure of precision, and provide gradations of progress that might serve as an incentive to motivate students. It also makes it possible for grades to be combined to produce overall grades (in ways that appear to exhibit quantitative robustness) by assigning percentage values to the grades, although these values can differ from place to place and particularly from country to country.

Importantly though, the meanings of these grades are given less by their percentage equivalent and more by where they are seen to place a performance in relation to what is considered acceptable or not by students and assessors. Elbow (1997) pointed out that these perceptions are not similar, students (in the USA) often viewing as unsatisfactory a B grade that assessors view as representing strong competent performance. (A similar attitude towards the unacceptability of a C grade can be found in students in the UK). O'Donovan, Price and Rust (2004) made a similar point, and noted that the meaning of a grade depends on its use and its context in terms of course, subject and the performance of other students. Furthermore, Elbow noted the (often not explicit) linking of the grade to its significance or the stakes involved, and that some

grades are more significant than others – the B (in the USA) and C (in the UK) grades described above, for example. His point was that many of the grades on which assessors expend their energies have little meaning (the difference between C- and C, for example), while others carry greater significance. He therefore suggested that fewer grades might serve the same purpose currently served by over 13, and that these might present the assessor with a much simpler discrimination task. He described this as minimal grading, involving scales with two (pass/fail, satisfactory/no credit), three (strong/satisfactory/weak, excellent/satisfactory/unsatisfactory) or perhaps four levels.

Grades as measurements

Knight (2006) has offered some useful additional observations on the assumptions and uses made of the measurements employed in the assessment process. He made the point that because assessment judgement is observer-relative (judgements of human thought and actions are dependent on the context and circumstances within which the judgement takes place, as was discussed earlier), the notion of measurement arising out of this judgement ought to be treated with caution. He suggested firstly, that any judgement of understanding or skill expressed by another could only be a rough approximation of its value and likely to vary according to the perceptual context of the assessor. Secondly, as a consequence of this, the assumption that any numerical data arising from this judgement could be viewed as having interval or ratio qualities was illusory. Such data could only be ordinal or nominal. He referred to some aspects of educational assessment as 'quasi-measurement' (p. 438) where the numbers generated by the process (the percentage equivalents of grades, for example) were descriptors rather than true numbers capable of being used in arithmetical calculations. These measurements might arise out of, or possibly permit, the ranking of different performances. He suggested that their ordinal properties are a consequence of the non-determinacy of the qualities being assessed and the complexity introduced by the interactions of these qualities as they contribute to the assessor's overall impression of the performance.

Rust (2007) has expanded on the difficulties inherent in the ways assessment measures are used, noting that the problem is particularly endemic in higher education. In addition to the points made by Knight (2006), Rust argued that human assessors lacked

the capability to integrate systematically the number of cues involved in a typical assessment task and to make the number of discriminations implied by the assessment scales that are used. Given the importance of context, discussed previously, for giving meaning to assessment decisions (e.g. Shay, 2004, 2005; Scott, 2005; Knight, 2006), he suggested that it would be unlikely that the meaning of a particular assessment decision would be similar across assessments, assessors, subject areas, courses and universities. Because of all these difficulties, Rust (2011) has suggested that assessment judgements ought not to be expressed in numbers or conceptualised in statistical ways. The meanings of percentages are imprecise (but convey an illusion of precision) and they are not scalable, and therefore they are not amenable to arithmetic manipulation. The notion of a normal distribution should be thought of as relating to random events rather than the results of purposeful educational interventions. The nature of different assessment tasks and contexts, and different subject areas, additionally conspire against a commonality of meaning that might be attributed to these assessment measurements, and they are further contaminated by the inclusion of extraneous elements unrelated to academic achievement (Sadler, 2010).

These difficulties with the nature and use of the assessment criteria and grade or percentage scales suggest that it may be useful to consider alternative ways of conceptualising the scales. A possible extension of Knight's (2006) idea of non-determinacy coupled with a measurement scale consisting of multiple levels (e.g. 0-100%) might be to think of the measurements as consisting of ranges of values rather than a specific point on a numerical scale, and the widths of the ranges as being representative of the extent of the non-determinacy of the measurement. In a sense this already occurs in that grade and sub-grade bands are typically assigned certain ranges on the percentage scale to permit transformation of the quasi-measurement of grades into the numbers that are (inappropriately) used in further calculations. Additionally these ranges do not all have the same widths (30% for the A band, 40% for the Fail band and 10% for the rest, for example). However, the important point is that the non-determinacy of the measurement suggests that it may be useful to think of the measurements encapsulated in grade-bands in more flexible ways, as consisting of ranges of values having varying widths that are positioned along the scale at varying

intervals and potentially overlapping with adjacent ranges to varying extents. The positioning and sizes of these ranges may also vary from assessment to assessment and this variation may depend on the application of some of the criteria discussed above.

Summary

In this section on the nature and use of criteria a number of aspects have been discussed. The first is that there are two aspects of meaning encapsulated in the criteria, which relate to the writing and the system of grading, and two ways of conceptualising the writing, as a procedural skill or a means of communication. These two ways of conceptualising writing result in two broad types of criteria, those relating to the surface presentation of the writing and those relating to the content. The content consists of the information that is presented and the way in which it is presented (the analysis, reasoning, logic or argument, which may be representative of the writer's thinking about the information and the purpose behind the writing task). These aspects of writing have been conceptualised as emerging from a relatively homogeneous academic culture as a result of socialisation, where the writing is viewed as being relatively independent of the writer and reader. Alternatively they may be representative of heterogeneous academic contexts in which there are competing interpretations of the purpose and form of the writing (an academic literacies perspective). Although assessors, particularly in higher education, describe themselves as valuing the communication aspect of writing, they often focus on the procedural or surface aspects of it in their comments.

A second aspect of assessment criteria is that they may be more complex than is suggested by their surface specification, or they may involve additional meanings that are influenced or triggered by other variables. These relate to characteristics of assessors and views they hold about assessment and the broader education context. Thus assessors may bring additional, idiosyncratic or implicit criteria to bear on their assessment which may modify or become blended with explicit criteria, or explicit formal criteria may be ignored. Additionally the meanings of words used to specify criteria may be different in different contexts and when used by different assessors. Assessors may be unaware of these variations, suggesting that there are aspects of assessment that may be unconscious as well as private. This variability suggests that

criteria may not be well aligned with what it is that assessors do or need to do while assessing, and help to account for the low consistency of assessment across assessors and assessments.

Another variable that may influence the meanings and application of criteria may be the quality of the writing. Some characteristics relating to procedural or surface aspects of writing are phrased in negative terms, while those relating to content-based understanding are phrased more positively. Additionally some criteria are more commonly associated with poorer writing while different criteria are associated with better writing. Alternatively, criteria can be commented on in both positive and negative ways or be utilised in the evaluation of both poorer and better writing, relating more to complex and inter-related aspects of the writing that involve the writer's argument and thinking about the topic. These findings suggested that basic, less sophisticated writing may be more easily evaluated using discrete criteria. However better writing requires integration that makes it more difficult to consider some criteria independently of others. Analytical application of criteria may only be possible when the writing being evaluated is relatively simple. As it begins to express more complex ideas, the criteria used to evaluate the work become increasingly interconnected. Assessors begin to pay greater attention to the overall effect of the writing rather than specific discrete aspects of it. The matrix of meanings that constitute the assessment of a good piece of writing may not easily be captured by an aggregation of relatively static and unidimensional characteristics.

The third feature of criteria that was discussed related to the ways in which grades are used (or misused). Some grades are considered more significant than others and some of the distinctions made by assessors are more important than others. Studies suggested that criteria may not be viewed in isolation from each other and aggregated in some way, but may be used to arrive at a holistic impression of the writing. In part this may arise from the difficulty inherent in that aggregation, inviting the observation that an increasingly detailed specification of criteria may have the perverse effect of forcing assessors to disregard them because of their human computational limitations. This complexity may mitigate against assessor consistency and the likelihood of assessment

decisions having the same underlying meanings, and may help to explain the introduction of spurious factors in the assessment process. Given the observer-relative and non-determinate nature of assessment judgements, these may be considered at best as having ordinal qualities. For this reason the arithmetic manipulation of assessment results, whether as part of the process of arriving at specific judgements or when combining several such judgements to serve wider summative or classificatory ends, and the statistical inferences that arise from it, may be poorly founded and inappropriate.

Judgement

Introduction

The third aspect of assessment (in addition to meaning-making and criteria, as already discussed) that is largely private to the assessor involves how assessors arrive at their judgement of the worth of the writing and assign a particular grade or mark to it. They appear to do this by combining their interpretation of the writing, their understanding of the criteria against which it is evaluated and their construction of the grading or marking system they use to assign a value to how well the writing satisfies the criteria. It also appears to be the least researched aspect of assessment, and as a consequence the most obscure. Part of the reason for this may be the difficulty of gaining access to this feature of the assessor's thinking, and the possibility that it may be difficult to disentangle this aspect of assessment (in the assessor's mind) from the process of reading and the application of criteria and a grading system. Cumming, Kantor and Powers (2002) also suggested that the decision-making behaviours of assessors occur rapidly and concurrently, making it difficult to disaggregate them analytically. As has been suggested in earlier discussion, judgement may not be a separate process and some researchers (e.g. Milanovic et al., 1996) have not made this distinction. This may be another reason for the judgement aspect of assessment being less evident or overlooked in earlier studies. Others have made the distinction (e.g. Cumming et al., 2002; Crisp, 2008a) but have chosen to embed the judgement aspect of assessment within the total process, such that elements of judgement were recorded as occurring but the manner of

their occurrence was not examined very closely.

The need for research into the nature of the judgement process has been expressed for some time (Vaughan, 1991; DeRemer, 1998; Huot, 2002; Lumley, 2002; Grainger, Purnell and Zipf, 2008). More recently research has endeavoured to examine this aspect and to explore ways of addressing the methodological difficulties of doing so. Suto and Greatorex (2008a, 2008b) and Crisp (2008a, 2010b) provided summaries of earlier work. A feature of this work has been that it has spanned a wide variety of different types and levels of assessment carried out in a variety of settings and focussing in varying degrees on the specifics of judgement. More recently researchers (Suto and Greatorex, 2008b and Crisp, 2008a) have also recognised that the judgements made by assessors may have similarities to other aspects of human decision-making. They have thus sought to incorporate into their research aspects of human judgement and decision-making that have been identified by Tversky and Kahneman (1974), Klein (1997) and Laming (2004).

Reading, interpretation, evaluation and judgement in assessment decision-making

Building on their earlier work, Cumming et al. (2002) described a framework that was refined down to 27 decision-making behaviours they had identified in assessors evaluating the writing of students using English as a second or foreign language or as a first language. Of these, eight were described as interpretation strategies and bore similarities to the processes of reading and meaning-making described above in the section on meaning-making. The remaining 19 were described as judgement strategies and were categorised in terms of three types of focus: how assessors organised their own judgement behaviour (self-monitoring); how writing was organised and ideas and arguments were presented; and use of language. With respect to the latter two types of focus, the framework provided a limited explication of these strategies and more a list of elements of the writing (or criteria, as discussed in the previous section) that assessors needed to take into account in their assessment. These judgement strategy items began with 'assess', 'consider' and infrequently 'rate' (e.g. ‘assess reasoning, logic or topic development’, ‘consider spelling or punctuation’ or ‘rate ideas or rhetoric’) without suggesting what might be involved in such assessment, consideration or rating.

In contrast to this the self-monitoring strategies provided a clearer indication of what an assessor might do. Examples were ‘compare with other compositions’, ‘define or revise own criteria’ and ‘articulate general impression’ (Cumming et al., 2002, p. 77). This lack of close examination of some judgement strategies is characteristic of earlier work and may in part suggest that certain aspects of judgement (e.g. self-monitoring) are more accessible than others. This in turn may suggest that some aspects of judgement, like criteria, exist closer to the surface of awareness or have more objective characteristics, while it is more difficult to gain access to strategies that are more tacit.

Cumming et al. (2002) developed their framework by analysing 'think-aloud' commentaries made by assessors verbalising their thoughts as they marked, following the procedure described by Ericsson & Simon (1993). Cumming et al. described a prototypical decision-making sequence that involved scanning the writing for surface features (length, format, paragraphs), interpretation (classifying language errors, identifying and interpreting content, relevance, coherence, topic development and organisation, and envisioning the writer's viewpoint), and arriving at a grading decision while summarising and rationalising the judgements leading to the decision. The interpretation in this sequence referred not only to discerning the writer's meaning but also the extent to which the writing satisfied the requirements of the task, which might be seen as an aspect of judgement. Thus the framework reflected a combination, or interweaving, or lack of clear distinction between aspects of reading and of judgement. This is probably not accidental, for Cumming et al. stressed that these decision-making behaviours were not isolated from each other but were ‘interdependent sequences of attention to variable facets of the compositions and to the assessment criteria’ (p. 73). They described how interpretations were gradually formulated into judgements of quality, and then of grade, while balancing different kinds of interpretations against each other.

The sequence described by Cumming et al. (2002) suggested that the process of judgement might involve two steps. The first was arriving at a sense of the value of the writing, which might be termed evaluation as it reflected something of a process. The second was assigning a grade to that value. This might more appropriately be termed

the judgement, as it reflects a decision point in the process. Cumming et al. noted that, as in other studies (e.g. Milanovic et al., 1996), assessors did not always follow one sequence. Some of their participants focussed sequentially on particular aspects of their interpretation or judgement strategies, whereas others interleaved their evaluation and judgement behaviours with each other. It is possible therefore that in some cases assessors cycle between evaluation and judgement, judgements being decisions reached that are then incorporated into further evaluation. Cumming et al. also noted that their participants tended to focus on the language aspects of the writing at the lower end of the marking range, and more on the ideas and organisation for good examples of writing, a pattern that was reported in studies discussed in the previous section on criteria. With writing graded in the middle of the range, their participants focussed on balancing positive and negative elements of both aspects, and they also expressed greater uncertainty and ambivalence in their judgements. Finally they reported that while some assessors marked each piece of writing sequentially and in isolation from other pieces of work, other assessors adopted what were termed macrostrategies - how they approached the overall assessment task. Their microstrategies, involving individual pieces of work, included a procedural aspect relating to the organisation of the task and a substantive aspect. The latter involved the assessors' interpretation, and sometimes modification, of the criteria in relation to the batch of essays they were assessing.

Cognitive aspects of decision-making

In 2005 researchers in the Research Programmes Unit at Cambridge Assessment began reporting on work aimed at achieving a better understanding of the cognitive aspects of assessors' judgement and decision-making. Part of the rationale for this was to identify factors that influence reliability and validity in assessment. A study by Suto and Greatorex (2008b) drew on the work on human decision-making of Tversky and Kahneman (1974) and more recently that reported by Gilovich, Griffin and Kahneman (2002) and Laming (2004). The work by Kahneman and his colleagues and subsequent researchers had as its basis the idea that there are two cognitive processes involved in human decision-making (Kahneman and Frederick, 2002). The one is intuitive (involving automatic, effortless, associative, rapid and parallel processing) while the

other is reasoned and reflective where the processing is controlled, effortful, deductive, slow and serial and involves the application of rules and algorithms. Theories utilising these two processes have come to be called dual-process theories or models (Sloman, 2002; Evans, 2007), and the two types of thinking have been termed System 1 and System 2 respectively (Stanovich and West, 2002).

System 1 thinking is sometimes viewed as a way of simplifying a decision (because it is unimportant, or too complex for System 2 processing) and hence reducing the demand on processing resources. Alternatively the choice of process may be elicited by the cognitive task and the context and framing within which it presents itself (Gilovich and Griffin, 2002). Evans (2003, 2006) suggested that System 1 and System 2 processes can be present concurrently, and that dual-process theories of reasoning therefore 'propose the presence of two minds in one brain' (Evans, 2003, p. 458). The automatic and generally unconscious or preconscious System 1 mind interjects its perceptions, interpretations or conclusions into conscious and serial System 2 processes. The contribution made by Tversky and Kahneman (1974) to subsequent research in judgement and decision-making was the suggestion that System 1 processing might help to explain why individuals make judgements that do not always accord with a rational computation of the variables involved in the decision. They suggested that heuristics or rules of thumb were employed in System 1 processing, introducing biases that resulted in decisions deviating from rational best solutions. In later work (e.g. Kahneman and Frederick, 2002) it was recognised that a common feature of these heuristics was attribute substitution, involving the substitution of easily accessible proxies for computationally complex attributes. So, for example, the 'availability' heuristic involves the use of information that easily comes to mind rather than less available information that is actually relevant to the decision. The 'representativeness' heuristic suggests that the probability of an object belonging to a particular category, or of a causal relationship between events, is judged on the basis of their similarity, or the extent to which one is representative of the other. These substitutions typically take place without the conscious awareness of the decision-maker, hence the introduction of bias into the decisions.

A distinction between judgements based on intuition versus deliberation has also been made in the context of expertise and expert decision-making. Dreyfus and Dreyfus (2005) put forward an influential model of the development of expertise where they argued that expertise is characterised by very rapid unreflective or intuitive responses to situational cues. Where deliberation is used by experts, they suggested that its purpose is to act on and improve intuition rather than replace it. They also suggested that rather than expertise being the development of more complex, sophisticated and integrated rules for making decisions, it is defined by the increasing absence of those rules. They described five hierarchical stages in the development of expertise. Decision-makers at a particular stage act on the basis of intuitive recognition of stage-specific features of the decision while relying on rules and algorithms to permit them to incorporate the features of higher stages. An expert is an individual for whom all aspects of the decision-making process are intuitive or automatic. Dreyfus and Dreyfus suggested that this type of decision-making is not a result of the rules or deliberation becoming automatic (e.g. Osman, 2004), but the evolving of a more developed form of decision-making. On the basis of extensive exposure to cues and an emotional involvement in the outcome of actions in response to those cues (which serves to develop a sensitivity to the discriminative relevance of cues) the expert is able to develop an intuitive recognition of multiple, subtle, complex and interacting cues that permits him or her to function at a level beyond deliberation and the application of rules. An earlier Recognition-Primed Decision (RPD) model described by Klein (1997) also incorporated, on the basis of prior experience, development of a capacity for the recognition or matching of cues and their consequences. An important feature of this model was that the expert needed experience in order to be able to recognise the salient features of situations and mentally simulate various actions to resolve the situation. In unusual or misidentified situations it is possible to get a failure of either the recognition or the simulation, resulting in an inappropriate decision.

The intuitive aspects of the recognition and mental simulation that are carried out by experts, or the cognitive processing that emerges as expertise is developed, might be classed as System 1 processing. Where these are insufficiently developed it is necessary for the decision-maker to resort to more deliberate System 2 processes. Thus the way in

which the processes are utilised may depend on the experience and expertise of the decision-maker. Both types of processes may also be needed for less experienced decision-makers to function. The development of decision-making from explicit deliberative System 2 to implicit associative System 1 processing also supports the idea that varieties of decision-making exist on a continuum (Osman, 2004). Finally, rather than System 1 processing being simple or more primitive, or a replacement for deliberative reasoned thinking (e.g. Evans, 2003, 2007), the work of Klein (1997) and Dreyfus and Dreyfus (2005) positions System 1 processing as developing from System 2 processing and as characterising more sophisticated levels of judgement.

Dual-process decision-making and writing assessment

Suto and Greatorex (2008a, 2008b) suggested that the process of marking written work might involve both System 1 and System 2 processing, and that simple matching might engage the former while the careful application of a marking scheme would involve the latter. As assessors become familiar with the marking of a particular assessment and its related marking scheme, they might shift from System 2 to System 1 processing. In a study which examined the cognitive marking strategies of GCSE examination assessors, they identified five strategies which they interpreted within a dual-processing theory of judgement. The five strategies identified from participants' 'think-aloud' commentaries were matching, scanning, evaluating, scrutinising and 'no response' (the very simple decision required when a script contained a blank answer). They characterised matching and scanning as involving very rapid and intuitive System 1 judgements. They also noted that there was very little in their participants' commentaries that illuminated the decision-making involved, especially when the student's response was correct. When it was incorrect or unexpected they suggested that additional attention and perhaps an additional marking strategy might be needed.

Evaluating and scrutinising were described by Suto and Greatorex (2008a, 2008b) as System 2 processing, as these were interpreted as being slower and more effortful than matching or scanning. They also observed that their participants' 'think aloud' commentaries were more elaborated with these strategies and provided evidence of self-awareness and reflection that offered a richer description of what the strategy might

involve. They suggested that these strategies were more likely to be used when assessing more difficult higher-level assessments, and when a student's response was unusual or sophisticated. They viewed evaluating as a strategy in which assessors considered and collated a number of dimensions of an answer, and evaluated its meaning in order to arrive at a judgement of its worth and the marks it should be allocated. They found that this was the most frequently used strategy. Scrutinising was described as following on from, or being used in conjunction with, the other strategies, often when an unexpected or incorrect response was encountered. They described this as an attempt to work out why the response was incorrect or whether it might be a valid alternative answer. They felt that this strategy aimed to better understand the writer's thinking or line of reasoning.

Although they did not suggest it, a re-examination of the data provided by Suto and Greatorex (2008b) to illustrate these strategies suggested that it may be possible to identify a number of behaviours that might make up or contribute to the strategies. Thus data described as evaluating and scrutinising might be interpreted to suggest the discerning or labelling of the quality or property of an element (something is "fine", p. 223; "weak", p. 224; "enough ... to warrant three marks", p. 224) and reasoning ("only a simple answer", p. 224; "focused on one issue", p. 224; "continuing to analyse", p. 224). Additionally, within the evaluating strategy there was comparing ("which ... doesn't fit", p. 223) and hypothesising and confirming ("that looks ok ... just to make sure", p. 224). Within the scrutinising strategy there were behaviours that might be regarded as evidence of meaning-making on the part of the assessors ("er, is there some measurements? Hard to tell, but it looks as if they've ... they've extended back to find the centre of enlargement", p. 225; "the student has done it by finding one percent and half percent; again like trial and improvement working backwards thing, so obviously no real grasp", p. 225). The feature that appeared to separate data interpreted as evaluating from scrutinising was that the evaluation behaviours were followed by a comment on the number of marks the assessor awarded, presumably arising out of the evaluating taking place through each of these behaviours. This suggested that scrutinising might need to be followed by a process of evaluation in order to arrive at a mark, and that in some cases it may be a necessary preparatory step for evaluation.

Suto and Greatorex (2008a) also suggested that scrutinising might include evaluating. Thus it might be suggested that the number of strategies employed by their participants may be greater than was identified by Suto and Greatorex, and that there may be a measure of interdependence within the System 1 and System 2 strategies, as well as between them as they suggested. This may support some of the ideas in this regard put forward by Evans (2003, 2006) and Gilovich and Griffin (2002).

Thinking processes while assessing

Crisp and Johnson (2007) used the annotations made by assessors while marking examination scripts to gain access to their thinking processes while assessing. The participants reported that they used annotations to justify their marks and communicate the reasoning behind these marks, and that these supported their judgement process. This was particularly the case when the decision was difficult (for example, on a grade boundary). The annotations permitting them to return to things about which they were uncertain, or as a kind of "talking to myself" or "way of thinking aloud" (p. 957), and they helped them to "structure their thinking while marking" (p. 958) and create "a visual map of the quality of answers" (p. 959). This simplified the making of comparisons between texts and Crisp and Johnson cited authors such as Laming (2004) to support the view that making comparisons is central to the process of making judgements. The results of their study might be interpreted to suggest that the judgement process is made up of components derived both from personal needs and experience as well as assessors' communities of practice, while also reflecting a recognition of responsibility to both in terms of how the assessment is conducted. Sanderson (2001) made a similar observation in relation to the findings of his study into the process of assessment of A-level writing.

Recognising the limited research into the psychological processes involved in assessment, Crisp (2008a, 2010a, 2010b) conducted a study using 'think aloud' commentaries to explore several aspects of assessors' cognitive behaviour and related variables while assessing A-level examination scripts. These related to the nature of their reading behaviours, their social, emotional and personal reactions, the nature of their evaluations, and differences in evaluation behaviours between different assessors.

In the study reported by Crisp (2008a) the data were coded using codes derived from reading the transcripts and revised on the basis of other coding schemes. The codes were grouped into seven categories. Two of these were 'language' and assessment 'task realisation', which corresponded to two of the four core assessment criteria identified by Elander et al. (2006) ('use of language', and 'structure') which were discussed previously in the section on criteria and in relation to the more superficial aspects of writing. Two others were 'personal responses' to the writing by assessors (positive and negative affect, amusement, etc.), and assessors' 'social perceptions' of the writer (their characteristics and ability, teaching they might have received, and so forth). A fifth category was labelled 'reading and understanding', which included the scanning (System 1) and scrutinising (System 2) strategies identified by Suto and Greatorex (2008a, 2008b). This category contained codes that might accord with the notion of meaning-making that I discussed as a possible process preceding the use of criteria and the process of judgement. It might also be possible to interpret the 'personal response' and 'social perception' categories of codes as part of this reading and understanding process, or meaning-making – a part of the 'pseudo interpersonal relationship(s)' reported by Suto and Greatorex (2008b, p. 230), or an interaction by the assessor with the writer behind the writing.

The final two categories were 'evaluating' and 'assigning marks', and some of the codes within these categories appeared to relate most closely to the process of judgement in assessment. These codes showed similarities to some of the behaviours identified by Suto and Greatorex (2008a, 2008b) as evidence of a cognitive evaluation strategy (System 2), and may offer insights into what makes up this strategy. With respect to 'evaluating', some of the codes reflected behaviours that might more easily be discerned in 'think aloud' data (positive, negative, neutral or uncertain evaluations), whereas others ('weighing up' and 'compares', Crisp, 2008a, p. 254) might be the observable correlates of more complex processes. In terms of developing an understanding of judgement in assessment, it might be useful to attempt to find out what might be going on when an assessor is 'weighing up' or 'comparing'. Some of the codes categorised under 'assigning marks' might be seen as reflecting the aspects of assessment that I discussed in terms of criteria in an earlier section. They included the identification by assessors of assessment

objectives (evidence of knowledge, understanding, etc.) and references to the mark scheme or assessor calibration meetings. Others were more procedural (totalling and checking the mark). However, other codes again referred to aspects of the assessment process that could provide an insight into some of the deeper or more complex components of that process. Examples of these were 'first indication of mark or level' and 'mark decision/confirms mark' (p. 254) - at what point in the reading of the essay, and how and why, this occurs may say something about the judgement process. Similarly, the substance of a 'discussion/review of mark/reassessment', and of 'comments on own leniency/severity' (p. 254), may offer insights into how the assessor has framed his or her judgement. A consideration of these aspects was not reported in the study.

In a later study, Crisp (2011) described two additional behaviours: 'planning and orientating' processes that took place at the start of an assessment, and 'concurrent evaluations' that took place alongside reading behaviours. She also proposed an additional model of 'underpinning judgement processes' (p. 18) that complemented the sequence of behaviours she identified in assessment. Within this she described an inner frame consisting of 'reading and comprehending' and 'evaluative judgements of quality'. The latter was made up of a consideration of features of the work/mark scheme, comparisons between pieces of work, and concurrent 'analytic' and 'configurational' processes (Sadler, 1989). Analytic processes involved identifying cues/criteria, evaluating their quality or quantity, and combining them into an overall evaluation. Configurational processes (System 1) described an assessment of the work as a whole and its subsequent verification in relation to criteria, some of which might be implicit.

Much of the treatment of the results in Crisp (2008a) was discussed in terms of the relative frequencies of occurrence of the codes and comparisons of these frequencies across assessors and types of examination question. This reflected the aims of the study and underlying concerns about reliability and assessor agreement, and Crisp (2008a) suggested that differences in code frequencies might provide an indication of areas in which examiner bias might be more likely. However, this approach limited exploration of the data to try to explain the possible processes involved. Crisp offered in passing

some interpretations relating to these processes, but acknowledged in her later study that there was a need to explain these in greater detail. She suggested that constructing a representation of the writer's meaning was an important aspect of the assessment process, and that there might be 'instant evaluations' accompanying the reading (described as 'concurrent evaluations' in Crisp, 2011) and a later 'more overall evaluation' that resulted in a mark. Her participants sometimes linked evaluations specifically to task and assessment objectives (analytic processes) and negative evaluations were often associated with aspects of task realisation (missing material, understanding of the task, relevance) and use of language. She noted that these negative evaluations tended to co-occur with behaviours that were identified with making a decision about a mark, suggesting that the negative qualities of a piece of writing were more prominent at this point in the evaluation process. She also noted that participants often utilised their overall impression (configurational processes), rather than the marking criteria, when evaluating longer pieces of writing, as was observed by Lumley (2002) and Bloxham et al. (2011). As was suggested by Greatorex and Suto (2006), Crisp found that variation across assessors (as measured by variable frequencies of codes) did not adversely affect assessor agreement, and that different assessors were flexible in their use of different cues when carrying out their assessments. Where there was lower assessor agreement, with two outlier assessors (Crisp, 2008a), these assessors had lower frequencies of reading behaviours, comparing scripts and positive evaluations.

An initial model of the assessment process

In a later paper Crisp (2010b) examined the sequences and patterns of the codes to develop a model of the assessment process. Although she described this as a model of the judgement process (which might be viewed as one component of assessment), it incorporated a sequence of phases that encompassed the entire assessment process. It consisted of three phases and two additional 'Prologue' and 'Epilogue' phases. The Prologue consisted of the assessor's thoughts before reading the piece of work, procedural aspects of ordering and selecting scripts, expectations arising from previously marked scripts, or reminders about what the writing ought to be about. Phase 1 followed and involved reading and understanding the writing. This might be

accompanied by concurrent (or 'instant') evaluations, social perceptions and personal responses, and Crisp also included behaviours coded as language and task realisation as part of this phase. As I suggested above, these might all be interpreted as being a part of the meaning-making involved in the reading process. So, for example, Crisp interpreted the interjection "yes that's good" (p. 9) as an evaluation carried out concurrently with a participant's reading and understanding. An alternative interpretation might be that the data reflected the assessor's perception that what was being read fitted with the mental representation of the writer's intention that was being constructed by the assessor (i.e. meaning-making). Whether the data is representative of evaluation, or meaning-making, will depend on what the assessor meant by the comment. Unfortunately the nature of the data ('think-aloud' commentary) makes this information unavailable to the researcher. Phase 2 consisted of an overall evaluation of the positive and negative aspects of the writing and how the evaluation might be quantified (and corresponded to her 'evaluating' category), followed by the Phase 3 mark decision which consisted chiefly of the behaviours categorised under 'assigning marks'. The Epilogue was described as thoughts occurring after the mark decision in which the assessor discussed, reviewed or re-assessed his or her mark, or perhaps justified the decision. Crisp noted that each phase might incorporate a number of behaviours or reactions, and that not all phases were apparent in the 'think aloud' data from each assessment event.

An interesting feature of the model was the extensive description of Phase 1 in contrast to the other phases. Crisp (2010b) incorporated into that phase five of the seven categories of codes she initially identified in her data (and partially included a sixth – 'evaluating'). She also made reference to behaviours characteristic of 'assigning marks' (the seventh category) when discussing concurrent evaluations accompanying reading and understanding, and suggested that concurrent evaluation might also involve deliberation and attempts to explain and make sense of the writer's response. Occasionally assessors might also make some overall evaluations (Phase 2) during Phase 1 reading, and Crisp discussed a feature of 'looping' (p. 13) where the assessor cycled between Phases 2 and 3. Taken together, these suggest that differences between the phases may be less distinct than implied by the model, and that assessors may not follow these categories of behaviours in a progressive sequence. If this is the case, then

it might also be justifiable to conceptualise behaviours in the Epilogue phase as part of 'evaluation' (Phase 2) or 'assigning a mark' (Phase 3) rather than something apart from these. Critical to this interpretation would be whether these behaviours led to a revision of the mark. Although it was not clear from the data reported whether or not this occurred, the description of the Epilogue phase suggested that there would not be much purpose in this phase if it did not. If the Epilogue behaviours might be seen as a part of Phases 2 and 3, then incorporation of these behaviours might help to develop an explanation of the evaluation and judgement processes. As I suggested above, access to the assessors' thinking beyond that revealed in the 'think-aloud' data might have contributed to this development. If the time sequence of the behaviours within each phase is less important, then the integration of cognitive processes across phases may be an important aspect of how assessors arrive at a final decision. Judgement in assessment may not be sequential and linear, and the imposition of such a conceptualisation may limit an understanding of the assessment process.

In spite of this apparent alternation between phases by the assessors, Crisp suggested that there was 'general consistency of the sequences of behaviours' (p. 13) that might permit inferences to be made about the cognitive processes involved in assessment. Limited evidence from the data was provided to support the reconstruction of processes involved in the evaluation phase (Phase 2), and the explanations of these were largely speculative. In part, this may have been because there was limited commentary on evaluative behaviours in the 'think-aloud' data, or that these aspects of assessment did not show themselves in such commentary. With respect to Phase 3 the data was more revealing. The way assessors arrived at a mark appeared to be rapid and automatic. This may have been an example of the assessors adopting a recognition-primed decision strategy (Klein, 1997), where the writing was identified as typical of a certain response for which the assessor already had a set judgement. Alternatively, the apparent rapid automatic decision may have arisen from how the data was organised into phases, or it may have been a consequence of the method of data collection. Aspects of arriving at a mark might actually have involved behaviours that were categorised as part of evaluation, possibly suggested by the 'looping' that was described as occurring between Phases 2 and 3, or the participants' commentaries did not reveal aspects of the process

other than the actual mark decision. Crisp also reported that the participants occasionally commented on the severity or leniency of their marks, which suggested that assigning the mark may involve placing it within the context of the broader assessment task.

'Think aloud' methods of data collection

From the discussion above it is apparent that to some extent the data available for analysis, and consequently the analysis itself, were dependent on how they were obtained. In several of the studies the authors have offered comments on 'think-aloud' protocols and on the limitations these place on data collection and analysis. Three main observations have been made about 'think-aloud' procedures. The first is that they may provide an incomplete record of the participant's thinking process (Crisp, 2008a, 2008b). This may be because System 1 processes are too rapid, automatic and less conscious, and System 2 processes place analytic demands on the individual that make it difficult to verbalise all aspects of the process while engaged in the thinking (Suto and Greatorex, 2008a). Additionally certain types of information (e.g. cues that permit the recognition of stimuli) may not be accessible to an assessor's conscious awareness (Crisp, 2008b, 2010b). The second is that thinking aloud may interrupt or interfere with the process, such that individuals find themselves restricted in the attention they are able to give to the task, or they need to stop verbalising their thinking (with the attendant loss of data) in order to be able to do it. This is more likely to occur with System 2 processing. Suto and Greatorex (2008a) suggested that the extent of the interference may depend on the individual assessor and on the type of cognitive strategy being employed, and they speculated on whether assessors ought to be required to 'think aloud' retrospectively rather than concurrently to reduce the possibility of interference. However this might also result in data loss. Thirdly, possibly related to its interference with thinking, the approach may additionally alter that thinking (Evans, 2007) thus compromising the validity of the data (Crisp, 2010a). Evans suggested that 'think-aloud' commentaries were more likely to reveal analytic than heuristic reasoning, or that individuals engaged in a task were more likely to talk about their analytical thinking – System 2 reasoning is more easily verbalised (Osman, 2004).

Crisp (2008b) conducted a study to compare silent marking with marking accompanied by a 'think-aloud' commentary. She found that the variation in the marks awarded was not greater than normal intra- and inter-assessor variation, and that her participants did not feel that their marking had been affected by the 'think-aloud' procedure. However, they did feel that the procedure was distracting and that they were consequently slower to complete the marking. Although the marks awarded were slightly stricter in the 'think-aloud' marking, this was not statistically significant. Crisp also found that the interview data she obtained from the participants did not provide much information about their decision-making processes while assessing. They tended to focus instead on the content criteria they used to frame their assessment. She suggested that improvements in the interview questions might have resulted in more informative data. She commented that the verbal protocols provided more information on the 'marking process' which, from the extensive analysis of Phase 1 reading and understanding (Crisp, 2008a, 2010b), probably referred to this aspect of the process rather than the evaluation and mark decision aspects of assessment. Cooksey, Freebody and Wyatt-Smith (2007) made similar comments about the usefulness of 'think-aloud' data for identifying the sorts of cues to which assessors pay attention and their limitations for informing an understanding of the judgement processes in assessment. This was similar to my own findings (Delaney, 2005). While I found that the 'think-aloud' data provided useful information on how the assessors interacted with and made sense of the writing, it was less illuminating in relation to the less superficial aspects of how they conducted their evaluation and particularly how they arrived at their final judgement. This suggested that it might be necessary to explore other sources of data in order to be able to gain access to these aspects of the assessment process.

Summary

Studies on the judgement aspects of assessment have identified at least three types of assessment behaviour: reading and understanding, evaluation (in which understanding is related to criteria to determine the value of the writing), and the mark decision. A common feature of these studies has been a predominance in the data of reading/sense-making behaviours, and a consequent focus on these in the analysis. There has typically been less explication of what assessors do in terms of evaluation and judgement. These

may be more difficult to isolate or less accessible to an assessor's conscious thinking. Reading and meaning-making have been considered part of the judgement process, or aspects of evaluation and judgement have been incorporated into meaning-making. I have suggested that in part this may be a result of the 'think-aloud' method of data collection. The use of the term 'interpretation' to describe both an assessor's reading/sense-making and her or his evaluation constitutes an ambiguity, and may help to account for reading being considered a part of judgement and justify the importance of not separating the two. I have argued that it may be useful to view reading/sense-making as part of the assessor's 'dialogue' with the writer while discerning the writer's intention. Thus the judgement process might be viewed as separate and consisting of two aspects: a process of evaluation, followed by a decision (the actual judgement). In addition to the possible theoretical advantages of doing so, the limited data pertaining to evaluation and judgement suggest that there may be differences in how assessors think about these and the extent to which they are able (or not) to make that thinking available to the researcher. There is a need for research to focus more specifically on what it is that assessors do when they gauge the value of a piece of writing and arrive at a grade decision, and on the possible different cognitive mechanisms and decision-making behaviours that enable them to do so.

Dual-process theories of decision-making may offer a theoretical framework for understanding evaluation and judgement in assessment. Perhaps because of the limited data that has focussed specifically on these, consideration of these processes has necessarily been speculative. Developing expertise in judgement may involve a transition from slow and effortful analytical System 2 processing to rapid, effortless automatic or intuitive System 1 processing, possibly as a consequence of increasing familiarity with the assessment task. Shifts back to System 2 while assessing may be necessary when an unusual or unexpected script is encountered. Alternatively, simpler System 1 processing may substitute for more complex System 2 processing, employing heuristics which may give rise to biases. System 1 processing may not be easily available to conscious awareness, thus making it less accessible for investigation. System 1 and System 2 processing may be representative of judgement and evaluation respectively, suggesting that they might occur in varying proportions depending on the

writing, the assessor or the assessor's progress with the assessment task, or they may relate in more complex ways to evaluation and judgement. Assessors may employ both types of processing, possibly in order to process different types of assessment cues or to carry out different types of evaluation or judgement. Additionally the two types may be complementary or interdependent, and the output from one type of processing may contribute to the other.

Studies have suggested that grading decisions are often accompanied by a focus on the negative qualities of the writing and that negative evaluations are more often associated with the more peripheral structuring or task realisation aspects of the writing. This was also discussed in the section on criteria. While assessors might consider specific criteria in their assessments, they often utilise their overall impression when evaluating a piece of writing. Assessors' thinking while evaluating and formulating judgements includes reasoning and justification, non-linear attention to different aspects of the writing (particularly with difficult decisions), talking to themselves silently or through their annotations, holding a 'visual map' of features of samples of writing within the batch they are assessing, and relying on the context of other marks. Importantly, assessors do not all show evidence of the same processes, behaviours or strategies, and these processes appear to be partly a function of their socialisation into a community of practice and partly the result of personal idiosyncracies.

The findings relating to evaluation and judgement are more in the nature of an identification of possible features than providing a coherent model of the cognitive processes or strategies employed when assessing writing. They suggest that the notion of an evaluation and a judgement phase or stage following reading and meaning-making has some substance, but provide a limited explication of the content and, more importantly, the relationships between the processes that make up the stages, and the relationships of these to the outcome of an evaluation and the ultimate mark or grade decision. I have suggested that in part this may be a function of the complexity and depth of these processes, in part a function of the consequent difficulty of gaining access to them, and in part a function of the type of data collected thus far in an attempt to understand them. This suggests that it might be possible to make a contribution to a

better understanding of evaluation and judgement by trying to focus more specifically on these aspects of assessment, and by endeavouring to use an alternative approach to data collection that might offer some access to those aspects of assessment thinking that are not easily verbalised by assessors while in the process of assessment.

Chapter 3: Research Perspective, Methodology and Methods

Introduction

My initial thinking about my methodological orientation to the research was governed by my preconceptions concerning the nature of assessment, the focus of my research, and the nature of the data I thought I needed to collect to address my research focus. I thought it reasonable that the way assessors perform their assessment might be revealed by their understandings of the purpose and practice of student academic writing and of assessment more generally, and that it was likely that there were procedures they followed in order to carry out their reading/meaning-making, evaluation and final judgement. This orientation guided my conceptualisation of my approach and my choice of methods to carry it out. However, these evolved during the project as initial findings clarified my research focus and the nature of the data I was obtaining, which in turn permitted me to refine my methodological understanding and the methods I employed.

Because of the way I had conceptualised assessment, my research focus was on two aspects of assessors' experiences: their understandings of writing and assessment, and the procedures they followed to carry out their assessment. Importantly, I sought to understand from the assessor's perspective the aspects of assessment that are private or internal to the assessor. The only means of gaining access to assessors' thinking is through the descriptions and explanations that they might offer. One possible approach is through the verbalisations of their thoughts as they carry out their assessment. However this approach has limitations in terms of the aspects of the process to which it permits access, and there may be features of assessment that do not reveal themselves in assessors' verbalisation of their thinking. Therefore it was necessary to involve a process of interpretation and meaning-making to expose or construct a possible version(s) of what assessors do and how they do it.

The focus of the research had implications for the nature of the data. These would be

determined by participants' ability to gain access to their possibly tacit understandings of what and how they assess. Of the two kinds of data suggested by the research focus, the conceptual frameworks that assessors bring to the assessment task were relatively straightforward in terms of how they might be elicited and used to provide a description of their understandings. The second type of data was less clear in terms of how it might be elicited, and the constructions that might be placed on it to explain how assessors perform their assessment. The difficulty with these data is that the processes are cognitive and internal to the assessor, and inaccessible to the sorts of data capture available with observable behaviour. As they have very few behavioural correlates or outputs, it is not possible to measure outputs as a function of experimental inputs as a means of discerning the nature of the processes. Thus the nature of the data also has methodological implications, and these will now be addressed in the discussion which follows.

Epistemological considerations

Crotty (1998) suggested a framework for organising approaches to conducting research, distinguishing between the methods and the methodology that guides the choice and use of the methods, and assumptions about reality and what the research might reveal (the researcher's theoretical perspective and epistemology) that justify the choice of methodology and methods. At an epistemological level, many of the arguments advanced to justify a qualitative approach to research focus on the distinction between assumptions of objectivity in the case of quantitative approaches, and the recognition of subjectivity when data are treated qualitatively (e.g. Crotty, 1998; Lincoln and Guba, 2000). Quantitative methodologies emphasise prediction, control and measurement, while qualitatively-oriented methodologies focus more on discovery, description and meaning (Laverty, 2003). Thus the theoretical perspective of interpretivism is a reaction to the positivism that assumes an objective reality (Schwandt, 2000). Crotty described constructionism as offering an epistemological position that lies between objectivism and subjectivism and in some ways links the two. Citing the work of Heidegger, Merleau-Ponty and Brentano, he suggested that while things have an objective reality,

their meanings have to be constructed by human consciousness. Rather than meanings being created and imposed on reality (a subjectivist view), they are constructed and relate to that objective reality. Consciousness requires an object of that consciousness, just as the meaning of an object is shaped by that consciousness (Lyotard, 1991). This notion of intentionality, or relatedness, is embraced by a phenomenological theoretical perspective.

One important aspect of a constructionist view of knowledge is that different individuals might construct different meanings of objects or situations, and in that sense there is an element of subjectivity to the meanings that are constructed. This may result in meanings being coloured by experiences, expectations and possibly agendas (Crotty, 1998). However, the object or situation also exerts constraints on meanings and these will to some extent depend on the nature of the object or situation. A second important aspect is that while meanings might be individually constructed, the constructions are situated within a social context and are culturally constrained. The social experiences of the interpreter provide a framework for making sense of the world. It is the language, symbols and collective understandings of the interpreter's culture that provide the means for imposing structure and communicating this to others. Different cultures will construct different meanings of the same phenomena. Thus constructionism, sometimes referred to as social constructionism, also implies that there is a social element to meaning-making (Crotty, 1998).

Crotty (1998) was careful to point out variations in the way the term social constructionism is employed. While some theorists (e.g. Giddens, 1993) regarded it as referring to the social origins of only social realities, Crotty preferred to see it as a means for understanding all reality. The 'social' in social constructionism refers to how meaning is generated rather than the object of meaning-making. However he also recognised an individual perspective in meaning-making, particularly where meanings are being constructed about things that are less socially situated, or where investigators try to set aside their socially situated meanings. This construction of meaning is more individual and instrumental than the collective meaning-making of social constructionism. Crotty suggested that the term 'constructivism' might be used for

individual meaning-making, although he recognised the lack of consistency in the uses of these terms. There is also a blurring of the individual and the social in constructionist/constructivist approaches to knowledge and meaning (Cobb, 1994). Individual meaning-making may take place through interaction with others (Von Glasersfeld, 1989) while shared understanding and jointly constructed meanings permit individual sense-making (Rogoff, 1990).

If the assessment of written work is seen as something more than mechanistically matching elements of writing to criteria, then it involves a process of construction - it is itself an interpretive act (Shay, 2004) and an evaluative act (Huot and Perry, 2009). The object of study is more than a passive experience or a static perception but involves the action of discerning or constructing meaning (the sense and value of the writing). Additionally the action is cognitive and not behavioural, internal to the assessor and largely devoid of behavioural correlates. I aimed to investigate the internal subjective reality of assessors which is itself the product of their constructions, and they are limited in their ability to verbalise their tacit understandings of what it is they do. Therefore, my sense was that what I wished to discover was unlikely to be revealed through the descriptions of assessors. The limitations of findings obtained using 'think aloud' data collection techniques supported this inference. It was likely to require more active exploration and construction of interpretations arising out of participants' descriptions. By co-constructing these interpretations with my participants I would be taking a social constructionist approach to the elucidation of their constructivist understandings. Although there is a broader social and institutional context that exerts an influence on assessment, the assessment task is typically carried out individually where external influences may be diminished or they may be set aside. An important aspect of my research was to maintain a sensitivity to both the individual and the social, and to the possible effects of each on the other.

Theoretical perspective

The epistemological position I have described above outlines two kinds of constructions

which are relevant to my research: the constructivist interpretations of individual assessors, only partly related to the cultural milieu of assessment, and constructionist interpretations of these constructs negotiated between myself and my participants, which may recognise cultural influences. Within this interpretivist theoretical perspective there were two ways of viewing the data. The understandings of assessors constitute phenomena that are peculiar to their own particular sense-making of the constituents of assessment. Examples of these might be their conceptualisations of assessment, academic writing, academic competence, notions around the concepts of argument and coherence, grades, borderlines and so on. Beyond this, these phenomena might be explored from the perspectives of different assessors, taking into account the shared cultural and historical meanings that constitute a community of practice, and organising them so that their relationship to the whole might be explicated and the whole may contribute to a deeper understanding of the parts. The data may thus be viewed from both phenomenological and hermeneutic perspectives, interpretations of phenomena relating to assessment offering a fuller account of the wider phenomenon that is assessment.

In adopting a phenomenological stance towards an assessor's internal reality, my purpose was to be open to reinterpretation, or alternative interpretations, of assessors' understandings of assessment, and to limit the effects of presuppositions we might bring to our exploration. In doing so I wished to recognise the distinction made by Crotty (1998) between two approaches to phenomenological investigation found in the literature. One focusses on subjective experience, viewing phenomena as representing experience or something distilled from accounts of experience, and structures data collection and analysis so as to minimise the influence of the researcher's presuppositions on the description of that experience (e.g. Willig, 2001; Laverty, 2003; Denscombe, 2007). The alternative approach, suggested by Crotty, focusses on the phenomenon of interest rather than a participant's experiences of it. This involves an element of objectification of the phenomenon, recognising its existence independent of subjective experience. It excludes presuppositions by cultivating an awareness of these and examining critically how they might influence interpretations, and by seeking alternative interpretations. This latter orientation characterised my approach to the

research. It permitted me to move beyond a description of participants' experiences to develop an interpretation of how these phenomena might be used to explain the process of assessment. I will return below to a consideration of the notion of explanation and the implications of this for a qualitative approach to the collection and analysis of data.

In approaching the data from a hermeneutic theoretical perspective I was seeking to place myself and my participants outside the text of my exploration with them of their understandings of assessment. Again this enabled us to bring an element of objectivity to our exploration of phenomena and their possible meanings in relation to each other. From the position of our academic cultural situatedness, we were able to explore additional aspects of individual phenomenological constructs such as the influences of intentions, previous experiences, and thoughts and debates about writing and assessment. As suggested by Crotty (1998), a hermeneutic perspective implies an intention to go beyond the identification and explication of phenomena to considerations of what these constructs might mean, and therefore how they might be applied in the service of assessment. Because of my insider position in relation to the participants and the focus of the research, this interpretation was likely to involve a sharing of meanings and co-construction of understandings about assessment. I was trying to reach a position understood by all, as suggested by Gadamer's philosophical hermeneutics (Schmidt, 2006; Langdridge, 2007).

As part of my hermeneutic stance I recognised that some meanings relating to assessment might not be part of assessors' awareness, but might be uncovered through our co-construction and negotiation. Gadamer (1993) argued that the nature of things was revealed through shared understanding. Exploration of our partial or rudimentary understandings, or pre-understandings, permitted us to build on those understandings by using the parts we brought to illuminate the whole, just as our developing concept of the whole offered a context for a deeper understanding of the parts. This was consistent with the notion of the hermeneutic circle, as described by Young and Collin (1988) and Laverty (2003), for example. Thus rather than our construction of new meanings being limited by pre-existing understandings derived from prior experiences, these enabled our joint meaning-making. As argued by Gadamer, our past and present understandings

are bounded by constantly developing horizons. It is at the point where they overlap or fuse that the sense of one is brought to bear on the sense of the other, and deeper understanding is possible.

In his hermeneutic phenomenology, Ricoeur (1970, 1981) identified the meanings to be found in the fusion of horizons as exemplifying the hermeneutics of recollection or empathy that consist of the immediately apparent meanings of a text. However, they may be coloured by lack of critical awareness of cultural and historical influences (Moran, 2000). Ricoeur suggested that there may be meanings hidden beneath the surface of the text, perhaps beyond the intentions or understanding of the author, that may be revealed through demystification or a hermeneutic of suspicion. The implication of this is that it is necessary to move beyond immediate consciousness to discern the complete nature of a text, attributing new meanings to phenomena and synthesising new meanings from novel rearrangements of phenomena. I aspired to bring this critical dimension to my analysis to discover additional implications within the data. However, I recognised that in doing so I may have been manufacturing something beyond the data, and this may have been more representative of my own presuppositions than the meanings within the data. As discussed by Schwandt (2000) though, discerning something as being representative of something else is the essence of interpretation, and presuppositions are the means by which we are able to understand (Gadamer, 1993). It was important though to bring an awareness of this to my research process. In my discussion of methodology and methods I will give an account of how I attempted to do so.

Methodological considerations

Returning again to the purpose of my study, I was interested in what assessors think about assessment – their perceptions, views and constructs about it. In addition to this I was interested in how they do it. There was a contrast here between two types of data. The first was phenomenological, or my inclination was to frame it in phenomenological terms, and it appeared to be amenable to being approached from within the

phenomenological and hermeneutic theoretical perspectives described above. The second type of data had to do with processes, or sequences or matrices of events, where events precede or follow other events or interact or co-occur with other events. These processes involved more dynamic constructs or mechanisms than the phenomena I have outlined previously, and potentially described how constructs operate on, and are operated on by, other constructs. Examples of these were how assessors set about the tasks of reading, evaluating and grading a piece of written work, make sense of the writing and its assessment, compare, match, weight and differentiate between aspects of writing, assessment and grades, rationalise possible inconsistencies, and incorporate their various reflections into the process.

These two kinds of data created a tension for my methodological orientation. Regardless of theoretical perspective, qualitative enquiry tends to focus on meanings, on understanding rather than explanation (Schwandt, 2000). Shute, Torreano and Willis (2000) described this as symbolic knowledge. The findings of studies emphasise structure (Kvale, 1996) and typically the qualities of phenomena are expressed in noun-like terms. Less explicit is a consideration of process, or how people act or operate – procedural knowledge (Shute et al., 2000) or phenomena that have verb-like qualities. Where action is described in qualitative research, it has gerund-like connotations. Ways of operating are labelled but not explained. This feature of qualitative research may partly be explained by its focus on description rather than explanation and prediction, on finding things rather than speculating on their origins or causes or how they work. It focusses on the end-product of human sensing, perceiving or constructing (cognitively) rather than how a person does so. Cause and effect, explanation, and prediction are seen as the domain of positivist-oriented research where there is an assumption of objectivity, and cause-effect relationships exist regardless of the perspectives or constructions of the observer.

The epistemological assumptions and theoretical perspectives of qualitatively-directed enquiry support an examination of assessors' constructions of assessment from within both descriptive and interpretive or hermeneutic phenomenological traditions (Langdridge, 2007). However, they do not fit easily with the aim of examining how

assessors use their constructions in the assessment process. Two aspects of data relating to processes might place them outside typical qualitative analysis: the constructions I place on the data/processes might be beyond the constructions of participants and my interpretation may not be a part of the meanings they have of assessment, and in doing this I am according the data a measure of objectivity or apartness from the participant, thus making a distinction between the knower and the known. This contrasts with the phenomenological and hermeneutic view of meaning and understanding (Moustakas, 1994). This view of process data became additionally apparent as I tried to locate this data collection and analysis within a method tradition. Most qualitative methods emphasise the construction or co-construction of meaning in a social context by the participant and the researcher, with the possible exception of the methods of grounded theory described by Glaser (1992). The process I was investigating is cognitive and takes place within the individual. While there are social and contextual features that impinge upon this process, these are also mediated cognitively. Traditionally internal cognitive processes have been conceptualised deterministically and investigated with quantitative methods, proponents of qualitative methods seeing the admission of such processes as undermining their philosophical position (Crotty, 1998).

There are difficulties with qualitative methods that might be used to frame and analyse process data. I described possible difficulties with phenomenological or hermeneutic methods above, although I emphasised, in my discussion of theoretical perspectives, the possibilities for viewing data within this tradition with a measure of objectivity. Two methods which accord data an element of objectivity are ethology and grounded theory (Polit and Beck, 2006). While ethology is intended to document behaviour, I was interested in cognitive behaviour which is not amenable to direct observation. A possible analogue of this that could be analysed might be descriptions of what assessors think they do cognitively as they assess. Because of the tacit nature of this cognitive behaviour, I thought that such descriptions and interpretations might be limited or incomplete without additional discussion. Turning to grounded theory, an assumption of this method is that meaning is yielded up by data regardless of the perspective of the investigator (Glaser, 1992; Strauss and Corbin, 1998). Alternatively, although meanings may be viewed as constructions (Charmaz, 2006), her approach to grounded theory

nevertheless posits a close link between segments of data and the labels, codes or meanings the researcher attaches to them. I was aware that the data I wished to analyse would have been subjected to a degree of interpretive processing by thoughtful, self-aware and reflective individuals. It was also likely to be subjected to additional interpretation during elicitation, such that its meanings would not be easily captured by allocating and synthesising codes following the methods of grounded theory. I also did not feel that I would be sufficiently divorced from the data to be able to approach it in this way.

For these reasons I decided to remain within the hermeneutic phenomenological tradition in my approach to trying to understand the process of assessment. However, I recognised that some of this understanding might be removed from the understandings of the participants. The meanings I attached to processes might in part have emerged from constructions that I placed on the interpretations of my participants, viewed from the perspective of an overview of several separate accounts of assessing. The cognitive 'behaviour' that constituted the processes was elicited using phenomenological and hermeneutics based methods and assumptions, recognising that this data might have been 'objectified' in order to arrive at an interpretation of a process or processes to explain how assessment is carried out. This aspect of the analysis depended to a greater extent on the constructions that I placed on how the participants appeared to use their constructs to perform their assessment. My role shifted from being a co-constructor of meaning to one of examining the meanings thus constructed from a more objective point of view, in order to develop a model of how these constructs might function in an assessment process.

In adopting these positions with regard to the elicitation and analysis of the data and the sense that might be made of the findings, it was important to recognise that the meanings I arrived at in collaboration with my participants, and on my own, would offer only a limited sample of the possible constructions that might be made of the assessment process. My model of the process would be constrained by my participants' ability to recognise and verbalise their constructions, and by the meanings that we co-constructed about their constructions. It would also be limited by the extent of our

understanding of assessment, and by the limited number of individual views I would be able to explore within the constraints of the study. It would nevertheless provide some insights into assessment and how it is transacted, and any commonalities that I might identify in the constructs of my participants might be generally representative of what assessors do. I was aware that I am also an assessor, and that some of my experiences and understandings would be similar to those of my participants. This was likely to have an influence on both our jointly constructed meanings and those I constructed that extended beyond our mutual understandings. I will discuss how I endeavoured to address these aspects of my research in the section on methods below.

Methods

Rigour in qualitative research

An important aspect of the findings of research is that they should bear some meaningful relationship to that which they seek to explicate, that they are authentically representative, trustworthy and related to the ways individuals might construct their understandings (Lincoln and Guba, 2000). Several authors have offered suggestions for how this might be realised in relation to qualitative data (e.g. Lincoln and Guba, 1985; Stiles, 1993), while others have synthesised these into guidelines for examining and safeguarding the rigour of the methods (Elliott, Fischer and Rennie, 1999). Lincoln and Guba (2000) make the point that the trustworthiness of research findings are dependent not only on the rigour of the method, but also rigour relating to interpretation of the data and situating the findings and the study within a broader context.

In relation to the rigour of the method, Lincoln and Guba (1985) provided four principles for achieving trustworthiness that mirrored the concepts of internal validity, external validity, reliability and objectivity utilised in positivist research. These were credibility, transferability, dependability and confirmability. Stiles (1993) organised criteria for rigour into two groups, which reflected trustworthiness of observations and data (approximating reliability), and trustworthiness of interpretations and conclusions (representative of validity). The first group emphasised the importance of the

researcher's orientation/preconceptions and the context of the research and the research process, and engagement with the data, providing examples to illustrate its relationship to interpretations. These reflected auditing and reflexivity that support the dependability principle of Lincoln and Guba, and grounding of the interpretations in the data to demonstrate confirmability (Tobin and Begley, 2004). Criteria for trustworthiness of interpretations included agreement across data sources, researchers and participant checks, and transparency and coherence of interpretation, demonstrating credibility or the fit between interpretations and participants' understanding (Tobin and Begley, 2004). They also included the usefulness of interpretations for participants and for modifying the researcher's thinking about phenomena. This may bear some relationship to the fourth principle of transferability offered by Lincoln and Guba (1985). In my discussion of epistemology, theoretical perspective and methodology I provided support for the dependability of my findings. The description of methods which follows shows how I incorporated considerations of dependability and confirmability into my data collection and analysis. Confirmability, credibility and transferability are demonstrated in the presentation and discussion of the findings in Chapters 4 and 5.

Design

The practices that I wished to investigate were predominantly cognitive and internal to the assessor, and my means of access to them was through the expression of thoughts and thinking by my participants. Data generated contemporaneously with the practice of assessment might be obtained by asking participants to think aloud while they assess, following the technique described by Ericsson and Simon (1993), or by utilising a representation of their thinking such as the annotations that assessors make as they assess. Alternatively, assessors' practices might be explored outside the practice of assessment through their recollections and reconstructions of their practice. This might be carried out through interviews, or by means of participants' free descriptions or written responses to questions. The limitations in 'think aloud' techniques identified in Chapter 2 provided a rationale for adopting an alternative means of data collection, and my feeling was that assessors' comments and annotations might show similar limitations. While both techniques permit an element of objectivity to the data

collection, there was a likelihood that they would provide some indication of assessors' sense-making and the endpoint of their decision-making as opposed to an explication of how they carried out the process of moving from the one to the other. As previous work has suggested that this process is obscure and inaccessible, participants' unassisted descriptions or written responses to questions were also unlikely to be useful. As was discussed in the previous section, to gain some form of access to the process it was necessary to construct a version of that process as it was understood by each of the participants. The most authentic way of doing this was through discussion with each participant in a way that would permit me to be as certain as possible that I had understood what each meant, and that would provide a means for them to shape the meanings as we co-constructed them. For these reasons I chose an interview approach to the collection of the data.

A difficulty in relation to the use of interviews to explore decision-making is the possibility that participants may offer reasons for their decisions that make sense to them after the fact, rather than the thinking that took place during the process of making the decision (that might be revealed by something contemporaneous such as a 'think aloud' commentary). They may also feel a need to cast their decisions in a favourable light, thus emphasising some aspects of them more than other aspects. An important element of my approach to interviewing was to focus on what assessors thought they did while assessing, rather than on how or why they did it. I also asked my participants to think about their assessment generally, rather than to focus on specific instances. As they would be thinking about their practices across many instances of assessment, I reasoned that this would make it more difficult to think in relation to any specific outcome. This would enable them to focus more easily on aspects of their assessment process, rather than on the end result of the process. Thus my focus in interviewing was on participants' decision-making, and our jointly negotiated interpretations of what they did, rather than their decisions. I was assisted in this by the reflective orientation of my participants, their awareness of their own sense-making, and their capacity for applying a measure of objectivity and honesty to their recollections and reflections.

I conceptualised the interview data as a representation of the thinking, or cognitive

'behaviour', that constituted the assessment practice of each assessor. I then used these representations to construct an interpretation of how they might function in the assessment process. The purpose of my interactions with each participant was to build an understanding of their thinking and processing as they understood it (what they did), rather than involve them in constructing an understanding of the assessment process itself (how they did it). I chose to collect this data by means of individual rather than group interviews for two reasons. Firstly, as assessment is typically an individual activity and may display personal idiosyncracies, it was important to obtain representations of these that were as close as possible to participants' own understandings of their practice. While focus group interviews may provide insights into how participants develop and modify their understandings, they may alter their individual conceptualisations of these (Wilson, 1997). The second reason was that, because of its tacit nature, I anticipated that constructing a representation of what assessors do would require prolonged intensive discussion that focussed specifically on one individual participant at a time. In addition to time constraints, the context of a group interview was not well suited to this requirement.

Participants

One of my motivations for the study was to investigate assessment in my own educational context, but I recognised that my findings would have broader application were I to incorporate assessors from other subject areas. As approaches to assessment are constrained by educational context and subject area, I wished to obtain data from participants who worked in other subject areas which were sufficiently close to my own to be relevant. An additional ethical issue related to my choice of participants was the possibility of participants being identifiable as a consequence of the closeness of their link to myself. By broadening my pool of participants to include individuals from subject areas other than my own, and by not revealing the subject areas of individual participants, I was able to increase their anonymity.

Volunteers for the study were sought from experienced lecturers within the school of health sciences where I am employed, taking into account the criteria described above. Thus they were a strategically selected group (Davies, 2007). Seven participants were

interviewed, of which three were male and four female. Their ages ranged from 35-60 years and their experience of teaching and assessing in higher education ranged from 6-19 years. Two participants each were drawn from speech and language therapy, psychology and health and social care and one from nutrition and dietetics. My initial intention had been to involve three or four participants from at least four different subject areas, but because of the amount of data obtained from the first seven and the emergence of a meaningful structure from the analysis of their data, I decided not to collect further data. In addition to these participants, I also carried out a pilot interview with a new lecturer. This enabled me to evaluate the structure of my interview questions and to begin to develop my approach to the research interviews.

Context of assessment

The participants teach in a post-1992 university. The institutional guidelines on assessment marking have a brief section on the desirability of marking against agreed criteria, providing annotations to justify marks, preserving the anonymity of students, and moderating the marking. The bulk of the guidelines are then given over to a set of generic band descriptors, appropriate to Level 6 work, linked to degree classifications (rather than types of assessment), followed by a brief guidance in the use of the generic descriptors. The intention is that percentages rather than grades should be used in assessment, which are then matched to the bands that equate to the degree categories. The rationale for this is that it will promote use of the full range of possible marks and permit the making of finer distinctions between pieces of work. With the exception of the bottom two bands (0-19%), each 10% band has five types of descriptor that relate to knowledge and its understanding or application, presentation and communication, analysis and critical enquiry, research and scholarship, and critical evaluation and reflection. One of the principles covering the use of the descriptors is that discipline- and assessment-specific descriptors should be developed within schools and/or programmes, according to the principles underpinning the generic descriptors. The guidelines also suggest that pass descriptors should be expressed in positive terms, describing the characteristics of work in a band rather than what is lacking in relation to a higher band. In cases where work shows characteristics of more than one band, assessors are required to use their professional judgement.

Although the university has these broad guidelines and band descriptors relating to assessment, there are no mechanisms employed by the institution to monitor the use of them. They also require some interpretation by assessors in order to make them applicable to assessment at the lower Levels 4 and 5, assessment across different subject areas and types of assessment, and to particular assessments. From my insider experience through informal conversations with colleagues, the guidelines are not given much consideration by assessors within the school of health sciences and they do not exert a strong influence on assessment practices. There are variations across subject areas in how assessment is conducted and grading arrived at, permitted or encouraged by the guidelines, and flexibility in how the institutional guidelines are interpreted and applied is assumed and tolerated. Within subject areas there is some agreement on the approach to assessment, and each subject area has its own rubrics and feedback forms that are developed within the broader university guidelines. Typically descriptors are matched to grade bands though, rather than the 10% bands recommended by the guidelines. A consequence of the lack of a strong institutional influence on the practice of assessment is that staff express some uncertainty about how to go about it, and they are forced to make their own sense of the process. Typically this arises from interaction between individuals who work together within particular subject areas, thus resulting in the variations in practice across centres and similarities within them.

Ethical considerations

There were two important ethical concerns that arose from involving participants from my own work context: the possibility for views to be identified with a participant and the related issue of potentially discovering something about practices that could be viewed negatively by the university and have adverse consequences for participants. An additional consideration within my own subject area concerned my position as head of the centre. Although we discuss and review our practices as a staff group on a regular basis, the in-depth nature of the interview approach I was contemplating raised the possibility of my colleagues having to divulge aspects or uncertainties about their practice that could prejudice my attitudes towards them in the future. There was also the possibility of participants feeling coerced into participating and having to reveal

aspects of their practice that they would rather keep to themselves.

To address these considerations I focussed on three principles in my approaches to, and interactions with, my participants. I emphasised to them the voluntary nature of their participation, I tried to preserve their anonymity, and I provided them with control over how their data would be obtained and used. Permission to approach members of staff was obtained from the dean of the school and the heads of the three subject areas other than my own. I invited colleagues within the four subject areas to volunteer to participate if they were interested, after explaining the purpose of the study and what participation would involve. In two of the subject areas (my own and nutrition and dietetics) I provided the information to an assembled staff group. In the other two subject areas (psychology and health and social care) I approached staff members after they had been informed about the study by their heads of centre. Within my own centre and the school I obtained more volunteers than I subsequently interviewed and I also encountered colleagues who declined to participate, suggesting that participants volunteered freely and did not feel coerced into doing so. With respect to the participants in my own centre, the manner and content of their responses during the interviews was similar to that of the participants from other subject areas. I interpreted this as evidence that they placed sufficient trust in me to safeguard and not abuse their confidences that they were willing to talk as freely about their practice as other participants. The names of participants and the name and location of the university in which we work have not been used in this thesis, and participants were not identified by subject area. In using extracts from the data, I have tried to ensure that the wording does not provide an indication of the identity of the person providing the data.

With respect to participants' control over their data, several safeguards were put in place. They were encouraged to say as much or as little as they wished during the interview or to withdraw from the interview at any point if they felt uncomfortable or unwilling to say more. They were also informed that they could withdraw their data from the study subsequent to the interview. All participants were provided with a copy of the interview transcript, and my summary comments and initial interpretations, to review and check for accuracy. In a follow-up discussion they were encouraged to alter

or qualify anything where they felt they might have been misunderstood, misinterpreted or misrepresented. Prior to beginning the interviews participants were provided with this information, and I emphasised the voluntary nature of their participation and explained how I planned to ensure anonymity.

Data collection

Data collection was carried out by means of audio-recorded semi-structured interviews in which I introduced broad topic areas related to the assessment of written work. These covered the purposes of assessment and writing, the procedure or mechanics of assessment, construction of writing quality, meanings of grade-bands, borderlines and scaling in relation to judgements, and views relating to writing consistency. (See Appendix 1 for the list of topic areas and general probes). I explained the aim of my discussion with each participant in terms of gaining an understanding of the process of their assessment decision-making, rather than simply exploring their rationales for the types of decisions they might make. I asked my participants to think generally about their approaches to assessing discursive writing (essays and long-answer examination questions) produced mainly by undergraduate students, although at times they also referred to their marking of writing at a masters level. Although I encouraged them to describe examples that illustrated their practice, I expressed my aim as trying to identify commonalities across each participant's assessment practices rather than interrogating specific instances of assessment.

My participants' responses guided my questioning and the direction of our discussions. I responded to whatever seemed most salient to them, trying to enter their frames of reference and responding to aspects of the phenomena they were describing by commenting, questioning and summarising. I rephrased these until participants were in agreement with the constructions we were placing on their descriptions and reflections. I endeavoured to set aside my own constructions in order to focus on their constructions, help them formulate these and produce representations of their internal cognitive processes that were their own. This constituted a process of data generation arising out of questioning and dialogue with each participant, in the manner described by Von Eckartsberg (1986, as set out by Moustakas, 1994).

Initially I found this difficult. As a consequence of my experiences and interest I recognised the potential for me to utilise the interview context to confirm my own presuppositions and hypotheses about assessment. While this can be appropriate where the aim is to compare phenomena or explore their co-occurrence (Kvale, 1996), the purpose of my interviews was to obtain representations of the cognitive behaviours of my participants. I was thus constantly aware of the possibility of my questioning and summarising reflecting my own views rather than those of my participants. Because of this I became sensitised to two tendencies during my interviewing. The first of these was when a response was lacklustre or expressed weak agreement with something I had said (e.g. "yeah, maybe, yeah", "I guess so"). In analysing my early interviews it was apparent that my preceding comments had expressed one of my own views, and the participants had neither enthusiastically agreed with it (as it mirrored their own and was therefore a legitimate encapsulation of their understanding), or disagreed with it. The ability of participants to agree or disagree suggested that it was something they had thought about and represented an aspect of their constructions. In contrast, I took weak agreement to imply that they had not thought about it, that my comment had not facilitated their construction, and therefore that it did not represent something that was part of their understanding. The second tendency was when I became conscious of being focussed on what participants were saying, to the exclusion of thinking about my own constructions. This was particularly salient when I was aware that what they were describing was not something that I had previously thought about, and the thought uppermost in my mind was that I was discovering new data. The more that this occurred as the interviews progressed, the more I felt that I had no need of my own presuppositions, and I was able to set these aside in order to pursue the constructions of my participants.

While conducting my interviews I was conscious of the point made by Langdridge (2007), when discussing the work of Ricoeur, concerning the detachment of an interview text from the context within which it is obtained. It was important for me to establish understandings of the phenomena I was investigating, while in the process of interviewing, by constantly questioning and checking my participants' meanings. My

aim was to obtain, during the course of the interviews, as comprehensive and accurate an understanding as possible of what the participants thought they did. In addition to this I conducted follow-up interviews to explore further those aspects of the data of which I was not able to make sense, and to check interpretations made subsequent to the interviews where my own views might have coloured those of my participants. It was important to ensure that they agreed with my understanding of their descriptions and constructions of what it was they thought they did. I conceptualised the constructions as representations of the thinking, or cognitive 'behaviour', that made up their assessment practice. I then used the representations from all of the participants to construct my own model of how assessors might carry out the process of assessment. In doing so I was able to use my developing understanding of the whole to illuminate my understandings of the assessment decision-making of individual participants, while also using the variety of cognitive behaviours employed by my participants to construct an understanding of the process of assessment as a whole. This reciprocal relationship between the parts and the whole is described by researchers working within a hermeneutic framework as the hermeneutic circle (e.g. Laverty, 2003). I attempted, through my discussion, to enter into the mind-world of my participants during the interviews. Through this I hoped to bring that mind-world to bear on the interpretations I would subsequently make of their cognitive behaviours. I was assisted in this by the fact that aspects of their ideas were similar to my own and their culture was not foreign. I did not always need to gain an understanding of their world from the text of the interview as we already shared aspects of that world.

Method of Analysis

Within a hermeneutic approach, meaning arises from the relationship between the researcher and the interview text, recognising the role of the researcher in the construction of that meaning (Langdridge, 2007). My aim in the analysis was not to code the data, but to try to position the ideas, notions, concepts and themes within the data in relation to other such realisations of the understandings of the participants. I was seeking to replicate or model a structure for the data that might exist in the mind of the participants, rather than in the data. To do this it was necessary to draw on my own understandings of the phenomena, and on the understandings I acquired during the

interviews of my participants' understandings. During the course of my analysis, I was conscious at times that my interpretations did not always arise directly out of the data (although this also occurred). Rather, the data were acting as a stimulus for my own thinking about what might be going on. I viewed this as reflecting the notion of 'appropriation' described by Ricoeur (1981). I was also conscious that there were differences in how I conceptualised the data from different participants, reflecting multiple constructions as described by Laverty (2003) in relation to hermeneutic analysis, which I interpreted as evidence for the dependence of my thinking on the data rather than on my own presuppositions.

The interviews did not all follow the same course, as the participants focussed on those aspects of the assessment process that were meaningful to them. They were also not always able to fully articulate their understandings. Consequently there were gaps in the data, across and within participants. As part of my interpretive role I endeavoured to bridge these gaps by making connections between different aspects of participants' thinking, and by making use of the understandings of some participants to fill gaps in the understandings of others. To do this it was necessary to read and reread the interviews, moving between sections of data relating to individual participants and between the data of several participants, to refine the meanings I identified during the several stages of my analysis. This enabled me to move repeatedly between the universal and the particular (van Manen, 1990) as I built up my understanding of the participants' constructions of assessment and of what these might mean for a more general understanding of the assessment process. There was thus a fusion of the data and its context, and the contexts of the participants and of myself as researcher (Laverty, 2003).

The data was transcribed verbatim. A similar sequence of analysis was carried out with each interview. This began with an initial reading and re-reading of the data during which I sought to distil or summarise the conceptualisations I was able to identify in the data, being careful to ensure that the link between my summary phrase and the data was clear. This was followed by a second series of readings of each interview, in the latter stages of the analysis interleaved with the readings of other interviews, during which I

began to record interpretations of what I thought individual summary phrases might imply, both in relation to themselves and for an understanding of the broader themes or aspects of assessment that began to present themselves to me.

In order to explore the relationships of pieces of data to each other and the emerging overall framework that I was constructing of the assessment process, I used three ways to display the data during the analysis. The first involved entering the summary phrases (linked directly to the data) and interpretations (derived from the data) into a drawing program that permitted me to move them around and group them in different ways as possible meanings emerged. A printout of an example of one of these data analysis sheets is presented in Appendix 2. Each piece of data was identified as a summary (square box) or interpretation (rounded box), and a page and line reference to the source in the transcript was included. In addition, 'higher order' interpretations that emerged during my exploration of the data within each sheet were indicated by hatched rounded boxes. Lines were used to make links, and arrows used to indicate possible directions of influence.

A second approach to visualising the data, particularly with regard to criteria and their relationships to the grade scale, was to enter the data into a spreadsheet, again with page and line references to the transcript. Colours were used to preserve links made by participants between criteria they described themselves using at different points across the scale, or where my analysis suggested possible links, or where a criterion was seen as applying to a range of grades (see Appendix 3). Finally, in some cases the most useful way of capturing the meaning of an aspect of a participant's practice was to construct a narrative summary that pieced together in a meaningful way several pieces of data from the transcript that appeared to relate to each other to describe something. Examples of this can be found in Appendix 4. This process of analysis was done for each of the participants, permitting the data to be organised into the broader themes that began to emerge during the course of the analysis. The emerging structures for each of the themes were then compared across participants.

This enabled me to build a synthesis that might constitute a coherent whole, or sensible

tentative meanings (Kvale, 1996), for each of the themes and how they might relate to each other in the overall process. My aim throughout my analysis was to try to account for all the views expressed by my participants, incorporating everything substantive said by them into my account of the process of assessment. The resulting structures of the data and their synthesis are described in the chapter on the findings which follows.

Chapter 4: Findings

Introduction

Four major themes were identified in the participants' understandings of what they thought they might be doing when assessing students' written work. Two of these related to the participants' understanding of the nature and purpose(s) of firstly assessment and secondly academic writing. These had been anticipated, as they had been used to organise the broad areas of questioning covered in the interviews. The third area of questioning during the interviews sought to examine the procedure or procedures participants followed to arrive at an assessment of students' writing. The relevance of these areas of questioning had emerged from an examination of the literature and an analysis of what I thought was relevant to an exploration of how assessors arrive at their assessment.

However, early in the interviews it became apparent that the assessment procedure was more complex than could be described by a simple series of actions carried out by the assessor. This was evident from how the participants, particularly when trying to describe their assessment process or what they actually did, spoke in terms which I felt reflected two separate themes. I categorised these as *criteria* and *processes*. The data supporting the identification of these two themes emerged from the participants' descriptions of the more procedural aspects of their assessment, and my probing to try to elicit what it was that underlay this procedure. These themes provided a more fine-grained explanation of what might be going on.

By *criteria* I mean the elements that assessors look for in students' writing, and the determinations which they apply to their judgements of the worth of that writing. Some of these criteria are therefore the explicit or implicit criteria relating to the writing itself, some of which may appear in the assessment brief or the rubrics and marking guidelines that are applicable to the assessment. Other criteria relate more to the internal meanings assessors use in framing their understanding of assessment and their evaluation process.

These sometimes overlap with the assessment criteria but are understandably less explicit, often to the assessors as well. It is also not clear whether the criteria relating to the writing and those relating to the judgements of the writing can, or should be, separated into two sub-themes, as often they appeared to relate to both at the same time.

Processes refer to mini-procedures, operations, functionings or techniques. These are not suggested as sub-themes but as an attempt to convey a sense of the variety of 'workings', algorithms, heuristics or manipulations that assessors use or perform during the process of assessing student writing. In some ways these processes act on, or are informed by, the criteria – the criteria act as inputs to the processes, and the outputs of these operations can feed into or act on other processes. It is these processes that appear to describe aspects of the decision-making that support assessment.

The findings presented in this chapter relate to the two themes of criteria and processes. Initially I had anticipated that participants' perceptions of assessment and writing, the first two themes, would contribute to an understanding of how they carried out their assessment. However, early in the data collection and analysis it became apparent that the themes of criteria and processes emerging from participants' descriptions of their assessment procedures provided the most revealing account of how assessors operate. I therefore focussed on these themes in subsequent interviews. Where participants' perceptions of assessment or academic writing contributed to an interpretation of the criteria and processes they employed, these are included in the descriptions of those findings.

Part 1: Criteria

Introduction

Criteria that participants incorporated into their assessment broadly related to two dominant aspects of the assessment task: the grading system and characteristics of the writing itself (Table 4.1). These constituted frameworks of meaning into which participants mapped features or qualities they discerned in the writing. Data suggested

that participants had developed a number of constructs relating to grade-bands, and to writing which they thought displayed characteristics which were representative of those bands. These constructs are discussed below in relation to five aspects of the bands: the C grade-band; low versus high bands; writing spanning several bands; grade-bands versus sub-grades; and borderline grades.

Table 4.1 Aspects of criteria

The Grading System: Constructions of the grade-bands	C grade-band
	Low versus high grade-bands
	Writing spanning several bands
	Grade-bands versus sub-grades
	Borderline grades
Characteristics of the Writing: Content versus Argument	Content and argument in grading
	Dimensions of content and argument
	Structure of writing versus structure of argument

With respect to the writing itself, the participants viewed it as consisting of two features. One of these was the content of the writing, the information or 'facts' presented by the writer. The other related to how this information was organised and presented to the reader, often referred to as the 'argument'. This delineation between content and argument provided a second framework of meaning that the participants appeared to use when assessing, or which characterised different ways in which they approached their assessment or grading. This aspect of criteria will be explored under three headings. These are content and argument in grading, additional further dimensions of content and argument, and the 'structure' of writing versus the 'structure' of argument.

The C grade-band

Several participants commented on how they felt that the practice of evaluating academic work in terms of grade-bands influenced the ways they approached and

thought about the process of assessment. P2 described how when she felt a mark "emerging" she thought that she was casting these thoughts in grades rather than marks or percentages, because the meanings grades have for people are different to the meanings percentages have for them. Other participants also spoke about the constructions they placed on the grade-bands. There was agreement that the C band was seen as the midpoint, or average or modal point, in the marking range. P1 described the writing in this band as being "conventional, received", that the "argument is actually a conventional argument, it's what the handout told them to say". P2 described the C band as "bog standard average, not brilliant, not dreadful". P4 described it as a "pretty basic answer", "textbook", "competent". P1 made the additional observation that he thought assessors perceived their own students to be slightly above average, so they might see the mode for a sample of writing as being approximately 55-59%. P4 also saw the "mean grade" as being a C+, and that much of his marking decision-making was centred around the grades C, C+ and B-. It was apparent that the constructions the participants had for the C and B grade-bands were more elaborated than those they had for the A, D and F bands. Reasons offered were that the majority of pieces of student work fell into the former bands, and assessors consequently spent more time assessing work of this standard. Additionally, in deciding on an evaluation of writing in this range, assessors were required to think more carefully about distinguishing between the relatively large number of scripts grouped within this relatively narrow grade range.

Low versus high grade-bands

The meanings of the grade-bands discussed by P3 provided additional insights into the possible differences in the nature of the mid-range bands and those at the two extremes. Rather than talk about what she expected to find in pieces of writing typical of each grade-band, she described what she thought she was doing when evaluating a piece of writing within these bands. When talking about the C band, she focussed on the difficulty she experienced in distinguishing between students' work, how she needed to think clearly about her decisions, how she struggled to justify her mark, and how she felt that her decisions within this band were not taken with much confidence. In contrast to this, she felt that her decisions became easier when the pieces of writing fell

in the B or D bands, and easier still when in the A and F bands, although the transitions across the grade-bands could present difficulties. This latter aspect of determining the grade was something raised by all the participants and I will return to this below in a consideration of borderline grades.

Part of the reason for decisions being easier in the bands outside the C band was that there were progressively fewer scripts between which distinctions needed to be made. However, in addition to this, it became clear that for P3 the nature of the work being assessed at the two extremes of the grade range was different and that the evaluation criteria were therefore also different. Towards the lower extreme (the F grade-band) there was progressively less content for which credit might be awarded. As described by P3 the task of the assessor was made easy because "if something is missing it's easy not to give marks for it". The task of the assessor was simply to identify anything that might be relevant for which credit might be awarded. As the assessor was "searching for marks" little further evaluation was required.

Although the evaluation task may be simplified at the lower end of the marking range, P5 and P6 commented that there could be more demands made on the assessor in terms of making sense of the student's writing - "you ... have to keep revisiting it and trying to pick out the facts from the waffle" (P5). P6 commented that D and F graded work consumed a disproportionate amount of time for this reason, and that the writing could be difficult to understand because the concepts could be "muddled" or because the writing style could be opaque (e.g. poor grammar). P5 also felt that she would want to put effort into understanding the writing out of a sense of "empathising with the student". With writing being graded towards the lower end of the range P3 also acknowledged her tendency to be generous in awarding credit and to "give the student the benefit of the doubt". As a consequence of this it was relatively easy for a student to move up this lower range - "you look at what the student would have to do in order to go from 0%-20% and then ... how much more they'd have to do to get from 80-100% ... we probably were much more generous to those that had done real rubbish". All of these features of the assessment process steadily reduce in their effect as the quality of the piece of writing improves. As the mark moves up into the D and C bands, there is

an increase in the need to move beyond awarding credit merely for the presence of something relevant to the writing topic (which may be difficult to discern), towards forming a judgement of how apposite it might additionally be.

In contrast to this, P3 felt that at the higher extreme (the A grade-band) the writing was characterised by being increasingly complete. The assessment decision now had less to do with simply finding something relevant (in this grade-band typically everything is relevant), and much more to do with deciding how well the material addressed the topic. There was little ambiguity that needed to be given the benefit of the doubt. P5 commented that better writing required less thinking of the assessor, and P6 thought that reading A grade writing was quicker than that awarded B or C.

P3 said that her main concern when evaluating a piece of writing in the A band was that she was being too generous. She felt that much of her uncertainty arose from the difficulty of evaluating just how good it was. One reason she offered for this was that excellent writing was typically unique, which made it difficult to compare it with other pieces. P4 said of work at this level that "the way they've tackled it isn't the same as all the other students have, so it's novel in that sense". Even if there is a comparator of equal quality, it is likely that the approach taken within it will be different. The evaluation of the writing is thus increasingly subjective in this higher range, "what appeals to me the marker" (P1). As the awarding of a high mark might reflect an idiosyncratic preference on the part of the assessor, there might be an increasing reluctance to award such a mark. A second reason P3 offered was that with writing of this quality, she felt that she encountered limitations within herself concerning knowing what could still be improved or what was unimprovable. This contributed to an increasing "nervousness" about the grade that contributed to an increasing inertia towards awarding higher marks. There was increasing resistance to awarding marks as the writing got closer to the upper end of the marking range, and it was consequently increasingly difficult for a student to move up through this range of marks.

When assessing writing near the extremes of the grade range, P3 thought that she was less likely to enumerate the components contributing to a high grade. She said that she

might make a comment "that sums up the whole thing", and agreed with my suggestion that she might be making more of a global judgement of the writing rather than evaluating the contribution of individual components. Similarly at the low end of the grade range she said that "it's a lot more simple, it's a lot more straightforward". Again she agreed that she was making more of a global judgement. In this case her feedback would be much more extensive, to justify the grade to the student, "but it would be easier". Examples she gave identified omissions within the writing, which she felt were relatively easy to enumerate.

Writing spanning several grade-bands

An added difficulty in evaluation related to making sense of writing where some components were good and others poor. This could include good and poor components of the same type (good and poor referencing in the same essay which might be offset against each other almost numerically). Alternatively there could be components of different types (good referencing but poor paragraph structure). Between these extremes there might be variations in quality between sections of an essay ("she's done a good introduction and no conclusion"), or between within-paragraph and across-paragraph structure. Furthermore, rarely could the quality of any component be conceptualised in bipolar terms. Rather it existed on a continuum between poor and good, which further complicated the combining of such components.

When evaluating a piece of writing in the C grade-band P3 said that "the feedback I give on someone who's got a 55% is a lot more extensive *because I've had to think*" [my emphasis]. In discussing this with P3 we recognised that it became necessary to break the essay down into several components (of varying quality on a variety of continua), and then to try to weigh these up, or balance or offset them against each other. In the case of poor writing, there are fewer good components that need to be incorporated. As they also typically lie towards the lower end of the continuum there is less offsetting to do. Similarly, with a good piece of writing there will be fewer poor components and they are unlikely to be too far down their continua, simplifying their accommodation into the judgement. Thus the writing is more homogeneous at the extremes, permitting a more global judgement to be made. As the quality of the writing moves away from

the extremes towards the mid grades (typically C), the number of poor and good components becomes increasingly similar, increasing the complexity of integrating them.

Grade-bands and sub-grades

In talking about grade-bands P2 thought that letter grades had distinct meanings for both assessors and students. For example, she described students as seeing a C as "something you don't really want to get" and a B as "something you're satisfied with", whereas most would see a D as "fairly disastrous". She explained that letter grades "lump things (together) to send a signal" that is different for each grade. She felt that there was "a qualitative difference ... a distinction" between each grade-band. The move from one grade-band to another was described as a "big jump" (P5) and a "threshold being crossed" (P3). These impressions of the participants support an interpretation of the grade-bands as having a categorical dimension, and that there was a definite disjunction at the boundaries between the categories. Each grade-band thus marks a step change in the quality of the work rather than describing a range of points on a continuum, and the boundaries are the points at which the categories change.

Within each pass grade-band there are three sub-grades in the system used by my participants: C-, C and C+ within the C grade-band, for example. Although these were recognised by all the participants as having their own distinct identity, less importance was attached to them and to deciding which of the three sub-grades best described a piece of writing. The reason for this was that all three sub-grades sent broadly the same 'signal' about the piece of work. P5 described how "to start with you would reach a decision about what bracket [grade-band] it falls in and then you would be thinking more specifically within that". This lends support to the categorical identity of the grade-bands and the sub-grades permitting refinement within that grade-band identity.

Borderline grades

Comments made by participants suggested that they saw the borderlines between the grade-bands as having a distinct identity (in addition to defining a point of transition when deciding between two grade-bands). In the same way as she viewed grades as

"sending a signal", P2 described how she saw percentage marks not as points on an interval scale but also as sending a signal. Thus for example a mark of 60% was "just in" the B range, which sent a different signal to a mark of 59%. The mark reflected the feeling she had about the grade-band into which an essay should fall, rather than that there was a 1% difference between two essays, and she would choose the percentage to send the signal. She thus made a distinction between a piece of writing being "clearly in" (meaning a grade like D+ or C- that was not near a borderline) as opposed to "just in or out" of a grade-band. P1 described seeing a mark of 62% as being a "solid" B- as opposed to 60% being "on the cusp".

Writing which fitted between grade-bands displayed characteristics of both bands in roughly equal proportions, presenting the assessor with the dilemma that "if something is borderline ... you are thinking ok is this really worthy of the upper mark or not" (P5). Although the work fell into the borderline category (which had an identity for the assessor but no corresponding identity in the marking scheme), it was still necessary to determine which grade-band best fitted the work. For P7 the identity of the borderline related to it defining this area of decision difficulty, where at issue was the variation in quality of the writing across different marking criteria. The perceptual discreteness of the grade categories persisted, provided that all or most aspects of the work belonged to the same grade category. The borderline difficulty was precipitated by a mix of components of differing qualities that matched the criteria for different grade categories.

There were two approaches to handling the grading of this kind of work. The first was to shift from grades to thinking in percentages, and to award a mark of 59% or 60% rather than C+ or B-. Thus these assessors described the borderline grade as "B- at 60%", or less commonly "C+ at 59%", suggesting that the writing was qualitatively different from a piece of work that was "clearly in" the grade-band. Although the work was identified within the broad meaning of the grade-band (as 'satisfactory' or 'good', for example), the assessor was placing a restriction on the evaluation. This distinction supports the interpretation of a categorical identity for both the grade-bands and the borderlines between them. Further, the borderline 'category' consists of two categories, each of which sends a different signal: "just in" or "just out/under". 'In' and 'out' refer to

the upper of the two grade-bands, suggesting that the evaluation decision is taken in relation to the upper rather than the lower grade-band.

Two participants talked about the identity of the borderline, but ignored it when deciding on the final sub-grade for a piece of writing because they felt compelled to use only the sub-grades in the university grading scheme. However, they were conscious of the possible effect of the percent-equivalent of the sub-grade on a student's final mark for a number of assessments, which presented an obstacle to raising the mark to the higher grade. The approach they adopted was to keep a tally across pieces of work (a number of examination questions, for example) of instances where a borderline mark was given the higher sub-grade, in order to balance it by awarding the lower sub-grade for other occurences. This then would produce the same result overall of two marks awarded on the borderline. Thus the application of criteria when deciding on a grade or mark depended on more than just the merits of the work being evaluated.

Content versus argument

In addition to assessment towards the ends of the marking range appearing to involve different types of evaluation using different criteria, the substance of the writing was also described as undergoing a change as the work progressed in quality from a low grade to a high grade. This arose out of the participants viewing the writing as consisting broadly of two aspects. In part this perception may have arisen from the broad formal assessment criteria used by the participants. The first consisted of the content: the 'facts', information, research findings, points and issues raised in the literature – material readily available that would need to be replicated in the writing. The nature of the content in a piece of writing was recognised as being influenced by selection on the part of the writer, which in turn required an understanding of what was relevant or appropriate to the topic and the context of the writing. Typically the participants tended to talk about criteria relating to content in quantitative terms, couching this aspect of their evaluation in terms of the number of items of content present in the writing.

The second aspect of the student writing related to what was often called 'argument'.

This referred to how the writer used the content to exemplify, support and justify the ideas and interpretations they used to make their point. It involved the writer placing a structure on the content in such a way as to offer a meaning or meanings beyond that contained in the content itself, and which was the writer's own contribution to an understanding of the topic under consideration. Criteria related to argument tended to be talked about in more qualitative terms and included words like “depth”, “clarity”, “evaluative”, “originality”, “creativity”, “different perspective”, “novel”, “unusual”. Relevance in this context was evaluated in terms of whether the content was related to the point being made by the writer, rather than just being not irrelevant to the broader topic as was the case with content aspects of the writing.

What might be seen in the distinction between 'content' and 'argument' is a possible distinction between the writing and the writer. As described by the participants, content is something that already exists, can be found in the literature, has a fairly fixed identity and is subject to selection by the writer for inclusion in his or her writing. It is a property of the writing and consequently has a more objective identity – it is susceptible to being counted, as suggested above. In contrast, the argument is a property of the writer and the writing is the manifestation of the writer's thinking about the topic. What is seen in the writing is a step removed from the event and requires an element of interpretation on the part of the reader, hence it being evaluated more in qualitative terms.

Content and argument in grading

For most of the participants, the focus of evaluation in the lower half of the grade range was on content while the presence of argument defined writing in the upper half. Thus content without any argument was unlikely to move beyond the C band, C or possibly C+ being awarded provided that all “conventional” content was present, it was “reasonably accurate” and there were “no mess ups” (P4). P1 viewed the desirable structure of a typical essay as beginning with the “philosophical issues and definitions” followed by the “classical research”. This part of the essay was the “historical precis” and it should “write itself”. P4 described this as an acknowledgement of the “classic, expected, mainstream body of knowledge” which provided a grounding for any

following argument.

P2 described an essay using the metaphor of a journey. It started by saying "you've given me this topic", then explained what the writer had read or knew, then described where the writer had got to and what he or she thought might be the case and what might be worth looking at, and then put forward a view based on the arguments that had been presented. She suggested that "anyone can do" the first bit, but then it is necessary to build an argument until the end point is reached and the reader can see the path that has led to that point. She saw this latter 'thinking' part as more complex and harder. P4 suggested that content might continue to build as the piece of writing moved up through the B band, but that it would show "increasing depth or sophistication" provided by extra material from wider reading beyond the reading list. In addition to this he referred to an "increasing clarity and originality of argument, thought process" and "different perspective" as a piece of work moved up from B+ into the A range of sub-grades. P7 referred to this as having "gone beyond, over and above ... the regurgitation" and thought that work in the B to A range included extra content (beyond the lecture notes and reading list) and that it had been handled in a way that showed critical thinking and that the student had "learned and understood what they were meant to do".

P6 expressed the change in writing (beyond replication of content) in terms of synthesis and integration. He felt that it was the use of evidence to inform an argument that distinguished writing in the B band. P1 described how if a piece of writing is good – it has all the appropriate expected content and has followed the typical essay sequence (the predictable bit) – then he is "waiting to see [it] ... move up a gear" to B. This occurs when the writing shows "some hard critical evaluation and originality". P7 referred to this as the writer having "gone that extra step". P2 thought that it was important for a piece of writing to lead somewhere rather than being a succession of reports of what others have said. P1 described how some essays (those that do not move up a gear) "tend to drag on ... writing by absence". He felt that "they can't do the last bit so they drag out the early body and they end up with a sort of four line conclusion". He also felt that "you have very precise views ... as to what is a B and a C", which lends further support to the existence of a distinction between the grade range

up to C+ and that beyond B-.

Thus it appears that up to (and possibly slightly beyond) the C+ grade, the focus of the participants was on the content of a piece of writing, the top of the C grade-band reflecting relatively complete, albeit conventional, content. Moving down the grade range from C+ the writing is characterised by an increasing absence of elements of content. P1 suggested that rather than being characterised by criteria specific to the D band, the D band was characterised by the absence of C grade characteristics. This perception would support the notion that writing in the lower half of the grade range is evaluated more quantitatively in terms of the presence or absence of content rather than in terms of the quality of that content, and that decisions regarding the content become progressively easier as the 'quantity' of content diminishes. The assessor's attention is drawn to the absence of content, this absence becoming more marked as the work approaches the low end of the marking range.

Although there is some overlap of content considerations into the B grade-band, the defining characteristic of writing in the A and B bands is the presence of an 'argument'. In the A band this is increasingly described as “complete”, although towards the top of the A band assessors express increasing uncertainty in their ability to discern that completeness. In the B band the argument is less complete, or may be missing elements, but it is described evaluatively more in qualitative terms. As a piece of writing moves up through the B and A bands, it is described not so much in terms of having more elements to the argument but as being qualitatively better. The assessor's attention is drawn to the increasing coherence of the argument as the work approaches the high end of the marking range. P2 thought that in addition to this it was the expression of an original opinion or idea that “separates out the ... real A grade students from the A-/B+ types”.

Thus it might be appropriate broadly to conceive of the letter grade marking range, centred on the C to C+ grades, as consisting of a lower 'content' continuum characterised by 'amount' of content, and an upper 'argument' continuum characterised by 'quality' of argument, as illustrated in Figure 4.1. Furthermore, there appears to be a

disjunction between the type of evaluation carried out within each of these ranges. While aspects of (typically sophisticated or unusual) content might contribute to the evaluation of writing in the B and possibly A band, a piece of writing is unlikely to cross the C+ to B- borderline without at least some form of rudimentary argument. Similarly, a piece of writing containing some element of argument is less likely to be awarded a grade below this borderline. When P2 was asked whether she thought that she might evaluate good versus poor pieces of work differently, she did not think that her process of assessment was different but that what she was forced to evaluate was different. With a good piece of work she was dealing with "more sophisticated things", whereas with a poor piece of work she would be dealing with more basic aspects.

However, the transition from C+ to B- also assumes that the content of the writing is reasonably complete. A difficulty arises when this is not the case, and the assessor is required to make a decision based on an integration of elements of a piece of writing that individually warrant different grades. It is in the C grade-band, where the likelihood is greatest that this integration will involve both types of evaluation (perhaps requiring the reconciling of D characteristics with B characteristics), and where the diminishing content-based generosity and argument-based parsimony have reached their minimum, that this will be most difficult. This may be an additional contribution to the difficulty and uncertainty that assessors described themselves experiencing when evaluating writing in the C grade-band.

Further dimensions of content and argument

The ways in which the participants talked about content and argument suggested that these concepts might admit further refinement. The distinction I have made between content and argument was not always clear in the participants' commentaries. This was possibly because they had not or did not make this distinction, and at times they spoke of content and argument in the same breath. In attempting to provide a clearer interpretation of content and argument, it might be possible to view each of these as having two dimensions. With respect to content, at a simple level it refers to the facts, information, research findings, references, ideas, views and so forth that I described above, although even these exist on a continuum related to the ease with which their

presence can be discerned in a piece of writing. These are some of the 'elements' of writing. But beyond this, there is the issue of whether the elements chosen provide substance to the writing. P7 expressed this as "that they [the students] would know ... that's something, that it's an idea that can be used, and it can be used in different ways by different people who have different perceptions". In addition to the content elements being broadly relevant to the topic, their value for the deeper meanings to be expressed through the argument needs to be recognised and understood by the writer. I glossed over this above, when I suggested that at the lower end of the grade range evaluation related more to the presence of relevant components than to the quality of those components. However, their relevance introduces a qualitative dimension, and the better the piece of work the more important the meaningfulness of the content becomes for evaluating the work. It is this "increasing depth or sophistication" (P4) that may account for participants mentioning content as an element of their evaluation in the B range. At some level, the strength of an argument may rest on the choice of content upon which it is based.

With respect to the argument contained in the writing, the participants often spoke of this in relation to "understanding". Again this had two dimensions. Through the argument, writers convey their understanding of the material and the topic area (the content). In the case of undergraduate student writing, one of the purposes of assessment is for the assessor to gauge the extent to which the student has achieved this understanding. However, in addition to this, there is also the student's understanding of the particular writing task. P2 phrased this in terms of the introductory paragraph and the importance of it explaining the writer's approach and the rationale for it, and for this to make sense to the reader. Thus the student must not only address the topic but address the way he or she chooses to address the topic – in other words "provide a framework". She thought that the reader should know what to expect in the essay and be in a position to judge whether the writer had done what he or she intended to do. P7 phrased this as "they are demonstrating that they understand, that they have got meaning, that they had been out and ... understood what they were meant to do". In relation to very good pieces of work in the A band, the writing often demonstrates a uniqueness that arises from the student's approach to the topic, and the justification for

this, within the argument. The student has not only tackled the task in a particular way, but has rationalised the writing such that there is an argument in support of the argument, an explanation of the reason for the explanation. It may be this element of reflexiveness in the writing that characterises pieces of work that are given the highest grades.

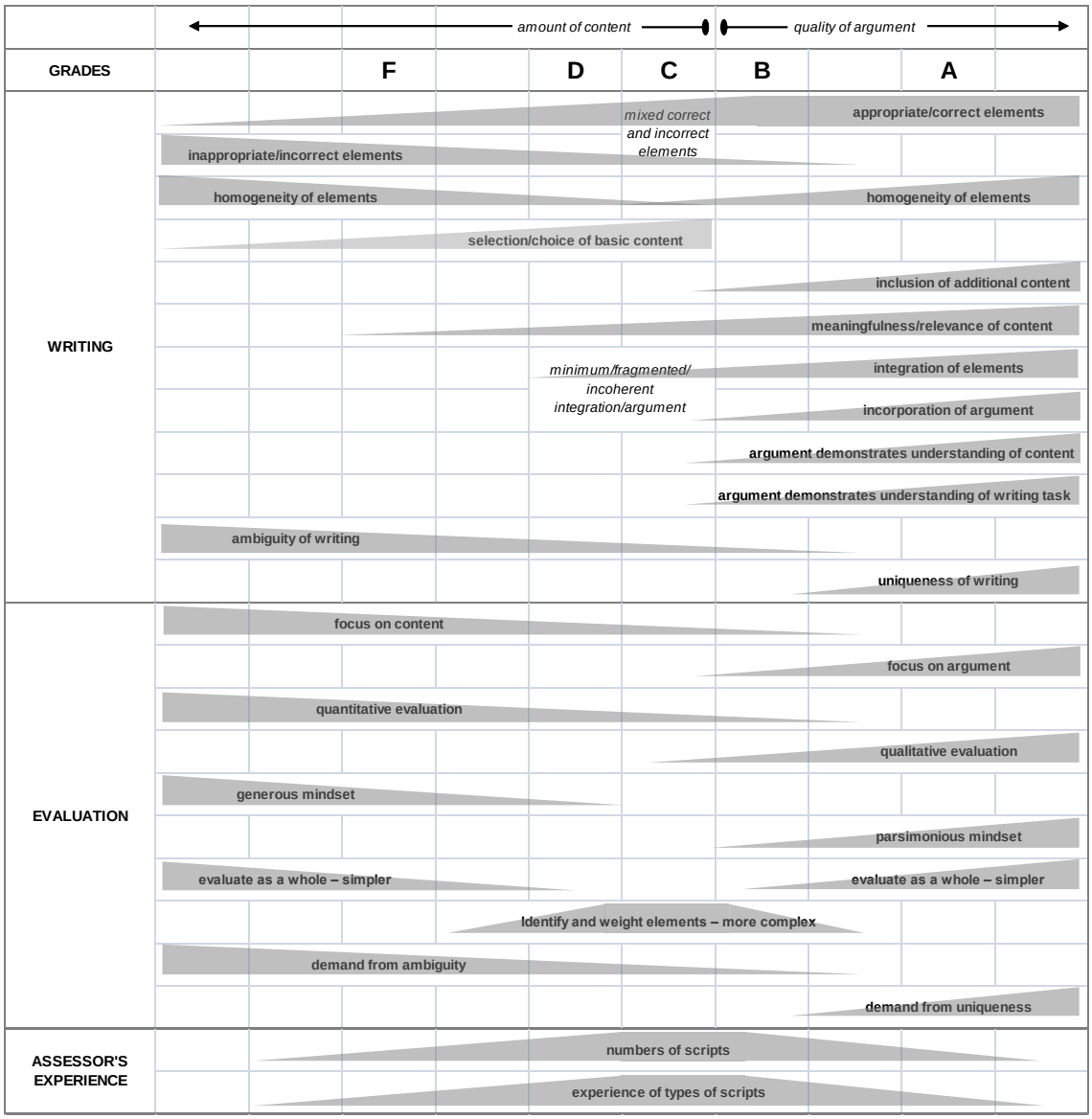
Structure of writing versus structure of argument

The participants were aware of there being a distinction between the structure or presentation of writing, and the structure or meaning of the argument presented within the writing. The former related to grammar, sentence clarity arising out of word order and sentence length, apposite use of words and terminology and so forth, while the latter demonstrated the writer's, or facilitated the assessor's, understanding. In spite of this they did not always make this distinction when talking about structure, and sometimes moved from one to the other while talking about either. For P5 there appeared to be a blurring of the expression of the writer's thinking (the argument), the logic with which that meaning was imparted ("structure and flow"), and the choice of words (or style) used to express that meaning.

Although presentation was viewed as less important than the meaning of the writing, P7 felt that one of the writer's tasks was to take the needs of the reader into account, and P4 thought that one of the purposes of writing was to engage the reader. Additionally, good presentation facilitated the reader's assimilation of the meaning of the text. P2 thought that it was difficult to lead someone through your thinking if you write badly. She thought that it might relate to the function of language in thought – if you can articulate things well you are probably also a clear thinker, and it means that you evaluate arguments well. The participants felt that presentation could only be awarded credit when there was sufficient content, and probably argument, to fulfil the main purpose of the writing. The more sophisticated the content (demonstrating the second dimension of content discussed above), and the more meaningful the argument, the greater the likelihood that good presentation would add to the writing. However, the participants were also aware of an element of linking or correlation between these constituents of the writing, and this may be why there was a tendency for participants not to make clear

distinctions between them.

Figure 4.1 Participants' constructions of grades and writing



Summary of Part 1

Part 1 of the findings has presented data in support of the idea that assessors make use of a number of criteria as part of their process of assessment. These criteria involve, or are related to, the requirements for the writing task, and assessment or marking criteria. Additionally, they extend beyond these to include the explicit and implicit meanings

that assessors have, both for grades or marks and the grading system, and for the nature and function of writing in an academic context. A feature of the criteria is that typically they only apply to a part of the grade scale, their values increasing or decreasing, and in some cases remaining constant, across the segment of the grade scale to which they apply. A summary of the criteria is provided in Figure 4.1 above, illustrating approximately how their values increase or decrease across the grade range.
The criteria provide a matrix of meanings that frame the assessment task and that assessors employ as they locate a piece of writing within the assessment grading system. The types of processes that assessors employ within this framework of meaning will be described in Part 2.

Part 2: Processes

Introduction

What follows is a description of the processes, or mechanisms, mini-procedures, operations, functionings or techniques identified from the accounts that participants gave of what it was they thought they did when they were assessing. They include basic algorithms, heuristics and manipulations as well as ways of capturing or representing features of the writing and the context and requirements within which it was produced and was being assessed. In addition to these processes, aspects of content related to the processes were also identified in the data. These were interpreted as components of the writing or direct and proxy markers or cues, in some respects related to criteria, that were used in conjunction with or were operated on by the processes.

An important aspect of the processes that was identified was that while some were common to more than one participant, none was common to all and several were represented in the account of only one participant. This may be a function of the data collection method and the direction in which the discussion moved while the interviews were being conducted. It may also be representative of those aspects of their assessment practice that were most salient or important, or most accessible, to the participants. Alternatively it might suggest that some assessors employ processes that

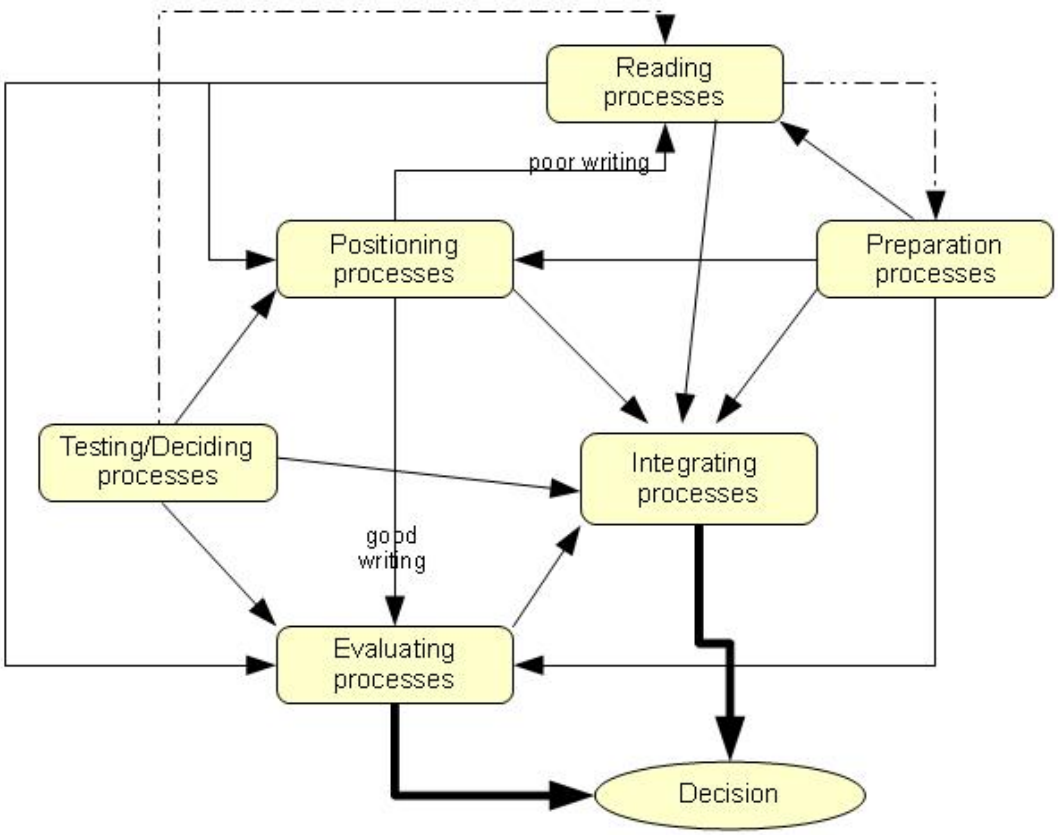
others do not.

Table 4.2 Assessment processes

Preparation processes	Identify expectations, form mental 'set', orientation; cultivate 'openness' - *create frame*
Reading processes	Scanning, identifying content elements, tallying, counting – *mechanistic and objective measurement*
Positioning processes	Maintain openness, suspend judgement; sensitive to 'trajectory' of writing; focus on the whole (supports evaluating) or on components (supports reading) – *subjective receptiveness*
Testing/Deciding processes	Identify something departing sufficiently from a criterion to be noticed; test for the presence or absence of a quality (rather than a component); make a choice/decision whether a criterion satisfied or a characteristic present or one component outweighs another; provide justification for a value – *objective application of conditions*
Evaluating processes	Emergence of sense of value, consideration of several components; establish a grade range then narrow range by comparison and matching components (within grade-band); testing and weighting components to determine whether a threshold requirement met (at a borderline) – *precedes subjective judgement of grade (uncomplicated assessment)*
Integrating processes	Consider profile, capture totality of the work; aggregate or integrate impressions and components, combine components (averaging, weighting, adding, subtracting, compounding, modifying); relate to broader purpose(s) of writing – *subjective synthesis resulting in grade (complicated assessment)*

Examination of the data that were thought to provide evidence of processes suggested that these might be representative of six categories of processes. They are summarised in Table 4.2 and will be described below. While it might be tempting to view these categories as representing stages, they may more usefully be viewed as modes of functioning while assessing. It is not always easy to see them as distinct, suggesting that they are to some extent inter-related, with each making use of aspects of the others, or interdigitating with them, as part of the processing. They also do not always appear to be used in the same order, and can be drawn upon at different times during assessment. Figure 4.2 provides a representation of the possible relationships that might exist between the six categories, with the arrows indicating the direction of influence of one process on another.

Figure 4.2 Relationships between process categories



When examining the data in order to identify and describe these processes, it became apparent that the participants often spoke about what they did in relation to elements or components of the writing they were assessing. Therefore, before examining the processes I will offer definitions that incorporate a possible distinction that might be made between these, and I will provide a summary of examples of these that occurred in the data.

Content elements and components

Content elements and components refer to the sorts of things that the participants described themselves recognising, identifying, tallying, weighting and so on while carrying out their assessments. They are not criteria, but rather markers or proxies that serve as indicators that map back on to assessment criteria. So, for example, appropriate referencing might be a criterion. Appropriateness in this case is, amongst others, a function of the choice and distribution and the number of references, the dates of publication, presence in the list of important authors, and so forth. Thus the assessor may use the number of references in the reference list, the distribution of references throughout a piece of writing (both of which can be ascertained at a glance), the registering of a date that is not sufficiently recent, and the recognition of a current author's name, as markers or proxies for adequate referencing by the writer. There may be several content element markers that relate to a particular criterion. Additionally, an assessor may not always read closely enough to determine whether the marker is actually relevant to the text within which it is located, or to the meaning with which it is associated, in which case the marker serves as a proxy for the criterion rather than a contributing marker for the criterion.

While examining these markers it became apparent that there might be two types. Those described above, and summarised in Table 4.3, are examples of what I have termed content *elements* and they appeared to be less substantial markers for criteria. In addition there are markers that appear to be more complex and substantial, and more difficult to describe or specify, that I have termed *components*. These are summarised in Table 4.4. Content elements are relatively easily specified and identified in writing, and appear to consist of a few simple characteristics that can be rapidly processed by an

assessor. They also appear to have a binary quality of being either present or absent, rather than having a quantitative value. Thus they appear to be features of writing that are amenable to rapid and immediate tallying and summing and have an accumulative effect on the assessor's perception of the work. Sufficient elements that are similar may eventually 'chunk' to form a more substantial component that is then handled by the assessor in a different way.

Table 4.3 Content elements

Proof reading	Presence or absence of obvious typographical, punctuation, spelling or grammatical errors
Introduction	Containing an indication of the writer's intention (regardless of whether that intention is worthwhile in relation to the topic)
Introduction	Containing a justification or rationale for the writer's approach (regardless of the inherent logic of the justification)
Reference list	Length of reference list may influence whether it is viewed positively or negatively
References, names, dates	Presence, regardless of their relevance or appropriateness
Appendixes	Presence is a marker or proxy for care and attention, or added depth to the consideration of the topic
Headings	Presence is a marker for organisation, regardless of relevance
Presentation	Delineation of paragraphs, use of white space, handwriting, etc. a marker for care and attention, or writer “knows how to write”
Points, ideas, concepts not presented in lectures or handouts	Presence, regardless of their relevance or appropriateness
Originality, unexpected content	Presence, regardless of their relevance or appropriateness
Inappropriate word choice	For example, use of the term 'proved' in relation to a finding or something contestable being presented as a fact

Table 4.4 Content components

"Conceptual stuff"	Theories, models, concepts – substance to the writing
Logical links, being "walked" or led through a sequence or argument	Within and across paragraphs
Marshalling of thoughts, putting arguments together	Organisation, arrangement and sequencing of points, explanation of relationships between points
Aim, purpose of writing realised	Sense of an overall point or conclusion being reached
Read around the topic	Range of literature brought to bear on the topic, possible introduction of relevant new or unexpected material
Absence of a reference list	Indication of a serious misunderstanding of a requirement of academic writing
Absence of an introduction	Indication of a lack of overall appreciation of the communicative function of academic writing, and of received practice
Serious date error	Indication of lack of awareness of the relationship of an idea to other ideas that preceded or followed it
References	Appropriate, providing support for points, ideas, arguments
Something that disrupts the reader's perception	Content that represents a significant departure from a general impression gained from previous content, positive or negative
Sweeping statements	Indication of a lack of awareness of contested ideas, possibly arising from limited reading
Introduction or "opening"	Provides a context or anchor for the entire piece of writing

Components are more substantial markers that may consist of an aggregation of elements, or may represent a more complex marker that is made up of a synthesis of less well-defined features that the participants described in less clear terms. They appear to be more likely to be recognised gradually, or for their identity to be formed gradually, and they are often held over for evaluation and integration in conjunction with other components. They maintain a separate identity and, unlike content elements, they are less likely to be subsumed in, or simply aggregated with, other components. They may also carry a lot of importance in terms of their significance for the quality of the writing because of what they imply. So, for example, P1 felt that a significant error with the date of a particular theory signified more than just the relatively superficial error, and suggested that it might cause him to question the theoretical basis of the student's understanding of the topic.

Preparation processes

Several participants appeared to go through a process of orienting themselves to the assessment, the task and the samples of writing, both in relation to the batch of scripts and the individual scripts. This began with an examination of the topic and instructions for the assessment, including any additional information that might have been given to the students. P4 described how he oriented himself to the topic ("I've got that topic ... I'm in that mode, it's working memory, it's perception, it's attention on whatever I'm marking ... I'm sitting in the subject matter"). P2 described herself tentatively deciding what she might expect from the students, and then scanning a sample of essays to see how the students might have addressed the topic. She might then make adjustments to her initial expectations, making reference again to the original instructions to ascertain whether the students' interpretations might legitimately be incorporated into the assessment. She reported that she approached the assessment with an orientation that her expectations would evolve during the assessment, and that she tried to cultivate an openness to the unexpected. She also emphasised that she did not have a rigid marking scheme, and implied that this too might evolve depending on what she encountered in the essays. P3 described her expectation of something evolving more in terms of the actual writing, having an expectation that the sense of the writing may evolve as she read it. Again this suggested an orientation of openness, rather than a more fixed sense of what she might encounter. Together, these data may support the interpretation that these assessors approach their assessment with the expectation that an aspect of what needs to be evaluated, and how this might be framed, will reveal itself during the course of the assessment process.

P5 spoke about her initial impression of a piece of writing ("within the first page you have set up an expectation as to what the rest of that piece of work is likely to be") in terms of how this oriented her to how she would need to approach its assessment, and the amount of care or effort she would need to devote to marking the script. If her initial impression was good, she felt she needed to read it with less attention, whereas a poor initial impression signalled a need to approach the writing with greater care. Her initial orientation also alerted her to a change in the work, by experiencing disappointment when good writing deteriorated, or surprise when poor writing

improved, and this might produce a shift in her impression of the work and the orientation with which she approached her subsequent reading of it. She described the emergence of this initial impression as unconscious - "I'm not sure it's a kind of a conscious thought, it's not like I stop after a page and think ... it's going to be a B, I think it's just ... a subconscious process that I go through". Unlike P5, P6 felt that "within the first five minutes I've got an impression of where I'm going to put that essay". However, he also referred to the way this initial impression influenced how he would read the remainder of the writing. With "the good ones ... I find that I can get through [them] fairly quickly, you realise people understand, they can express themselves ... usually you can see in the introduction whether people really are expressing themselves clearly and articulately, if they're not then those are the ones you tend to carry on reading ... and spending a lot of time underlining and making comments". These data suggest that these assessors, on the basis of their initial impressions, may approach different pieces of work with a different mindset. Given a point made earlier in Part 1 suggesting that poorer pieces of work might be approached quantitatively whereas better work may elicit a more qualitative approach, this initial orientation may also serve to place the assessor in a quantitative or qualitative 'mode' of assessment.

Reading processes

These processes were relatively easy to discern in the data and related to the behaviours the participants described themselves performing as they read the writing they were assessing. They described themselves as "looking for", "searching for", "scanning" or "identifying" content elements within the writing. P2 described herself as "monitoring for a logical sequence of links", noting when she needed to "go back ... to make the link" herself which "calibrat[ed] something in [her] mind as to roughly where they lie in terms of grading". This suggested that the process of reading may incorporate an ongoing shifting of assessors' orientation to the writing that may feed into their positioning when deciding on a grade. P5 related her scanning and identification to her initial impression of the work. She felt that, when this had been poor, she was conscious of looking for something to disconfirm this impression, but when it had been good, she sought confirmation of this and additional features of the writing that might

add to this impression. P3 also felt that she was more active in seeking out positive aspects of the writing, in order to improve a poor impression or confirm a good one. In contrast to this, P6 described himself as sampling the text until he felt that his initial impression had been either confirmed or contradicted, although he also demonstrated some predisposition towards seeking out positive aspects by describing his scanning as panning - "you see bits of gold or something in there". It appeared that for this participant his initial impression played a greater role in his approach to reading, serving as a reference point for his scanning.

Having found or identified content elements in the writing, the data suggested that some of the assessors maintained some kind of tally or count of these, and did so in a variety of ways. They appeared to approach this in an objective dispassionate way and there appeared to be a distinction in their descriptions between maintaining this tally and utilising the tally as a part of their evaluation. Thus evaluating processes appeared to act on the result of the tallying processes, rather than subsume them. In discussing this aspect of reading with P1, we identified a metaphor of a mental tally counter or "clicker", recording positive and negative tallies, that he thought might describe how he "held" these content elements in his memory as he read. P5 used a metaphor of buckets on a scale into which she thought she accumulated counts of good and bad elements. P1 described the clicker as recording a mix of positive and negative elements and how he "hovered" at the approximate average of these. From his description, it appeared that this means of holding the tallies worked while the positive and negative elements were relatively equally interspersed and the counts were not too disparate. However "when it goes very heavily in one ... or ... another direction" it could also produce a shift in his perception, which caused an adjustment in his overall orientation to the writing. (This relates to the positioning processes described below). P1 spoke of these aggregations in terms of "counting" them rather than tallying them, and of a "counter" as opposed to a clicker. He described his use of this as follows: "You've got another little counter keeping count of how many exceptions you might have to deal with at the end ... If you get to the end and you're very, very reliably hovered around mid C and nothing else has happened then it's a mid C. If you've got a couple of cautions come up and you're still at mid C you just give it a re-read". I interpreted this to suggest that rather than an

averaged tally of instances of positive and negative elements, sufficient of either would result in a form of critical mass being reached where their effect on each other was more compounding or multiplicative. To a certain extent it is possible that this critical mass effect might be less likely when positive and negative elements occur alternately, suggesting that the order in which the elements occur may influence the weighting given to them. These data appeared to provide evidence for the existence of two types of tallying or counting which functioned differently as part of the reading process. The tallying may be carried out on easily identified elements in a relatively simplistic quantitative way. The counting may relate to more substantial and integrated components. Either these may form part of an evaluation process (assessing the degree of fit between impressions and the writing), or they may exert an influence on that impression or orientation (positioning) in relation to the assessment. This may contribute to a more qualitative view of the writing

Positioning processes

This category of processes related to things that the participants described themselves doing throughout their assessment, or which marked a shift or change in the assessor's approach. They also appeared to fall somewhere between the more mechanistic and relatively objective reading processes and the more deliberate and subjective evaluating processes. The first of these was referred to as "reserving" or "suspending judgement", a feeling that it was important not to be quick to come to a conclusion. Thus P2 described how if something did not make sense she would "wait for an explanation", and P1 in a situation where he might encounter a very good and original introduction said that "you've ... got a ... caution and you're ... saying I don't want to be haloed, halo effected into an unfairly high mark by this clever opening". P5 felt that she needed to suspend her judgement while reading through an essay in order to perform her evaluation at the end – she constantly tried to maintain a separation between her subjective impressions and her assessment. P7 described herself as reserving judgement to the middle to end of the essay "because you do have surprises, even up until the end".

In spite of this desire to maintain an open mind towards each piece of writing, some of the participants also appeared to be sensitive to "where I think [the writing is] going to

go " (P4), or to what P1 described as the "trajectory" of the writing. For P4 his impression became increasingly established, and he reported finding anything that contradicted this impression increasingly surprising, as he progressed through the work – the work would become "more difficult to retrieve ... or screw up". Several participants also thought that they might view writing that gradually improved more favourably than an essay that gradually deteriorated towards the end (suggesting either a recency effect or a recognition that the later stages of a piece of writing may be more demanding, as was discussed in Part 1). An important aspect of arriving at a sense of the trajectory was that it appeared to influence how the participants viewed the content elements and components of the writing. With better pieces of work, participants felt that they were able to position it more easily towards the beginning of the reading, as a consequence of which they paid attention to the whole and felt that the assessment was simpler. With poorer pieces of work it became necessary to compartmentalise and focus on the constituent elements and components, to be able to tally or count them in order to factor them into an evaluation. It also made the assessment process more complex and P2 and P3 described themselves making comments or margin notes in order to keep track of these. Thus there appeared to be a gradual rough positioning of the writing in terms of its quality, or a positioning of the assessor in relation to how he or she viewed the writing. Additionally this appeared to occur earlier for better than for poorer pieces of work (where participants suspended their judgement for longer), and this then appeared to influence how the assessor subsequently approached the assessment of each piece of work.

Testing/deciding processes

Some aspects of assessment may involve specific mechanistic processes when a decision about something needs to be made. This may be the point in the assessment when the assessor needs to make a final judgement on a piece of work, but it may also occur prior to this point when it is necessary for a choice to made about a particular content element or component of the writing – whether it satisfies a specific criterion or a sufficient proportion of a more general criterion, whether it has a certain property or quality or that quality is absent, whether the value being placed on one or more of the components can be justified, or whether the value of one component outweighs the

value of another. Evidence for this type of process was not as extensive as for the other processes, and the interpretation was derived from the ways in which four participants spoke about how they dealt with some of the content elements during their assessments.

P2 described asking herself questions such as “does [this point] make sense?” or “do I need to adjust [an impression]?”, or deciding whether a content element (e.g. the reference list or appendix material) was “sensible“, “superficial” or “relevant”. While these data might be seen as the application of criteria, of interest for a consideration of process is how the criteria are being applied. One interpretation of the process is that the assessor may be matching characteristics of the writing to qualities that define 'sensible' or 'relevant', or noting an absence of characteristics that may suggest that the work is superficial, in order to assign it to a grade category. However, the data did not suggest that the assessor was relating this decision to a mark or grade range, or trying to narrow that range, as might occur during evaluation. Rather, the assessor appeared to be making a decision about her perception of an aspect of the writing, and whether that writing had reached or not reached a criterion threshold. The output of this decision might then be incorporated as a further component in the positioning or evaluating processes.

When making sense of something, or needing to adjust her perception, P2 used her own cognitive state to test an aspect of the writing. In deciding on relevance, the testing might be carried out in terms of more external forms of reference, although these might also involve aspects of subjective value. P4 described how “in a lot of cases ... maybe [writing] doesn’t jar, but occasionally something will jump out at you”. In discussing this, he agreed that in some cases writing “jarred” while in others things “jumped out”. Both of these occurrences represented departures from writing that did not jar, and he felt that his attention was being called to something negative or positive. P3 also described herself as becoming aware of disruption in writing or being “distracted by this sort of ridiculous spelling“ for example. She was also sensitive to when writers “just lose the flow”. P2 was aware when something was easy to read or well-written, and when the “reader has to do things” to make sense of the writing. P3 phrased this as “I have to keep re-reading sentences”. P1 described how he detected “hollowness” (lack

of content in otherwise well-written work) by becoming aware that he needed to read something twice, which "rang an alarm bell". Importantly these participants were not scanning for particularly negative aspects of the writing, but were conscious of reaching a point when they decided that these had occurred sufficiently for a decision to be made about them, such that they might be incorporated into an evaluation. P4 felt that this recognition of being aware of something might influence how he might begin to think about an aspect of the writing - "I might be starting to think yeah". This suggested that an awareness of departure from a steady state, or the lack of such a departure, might represent a decision point in the formulating of an impression – the assessor was deciding in terms of whether he or she was conscious of something occurring or not occurring.

Evaluating processes

Descriptions which were interpreted as being representative of evaluating processes were those that moved beyond tallying and positioning, and appeared to provide evidence of a sense of the value of the writing emerging, or being developed by the assessor. These processes involved the assigning of values to tallies and counts and the assessors becoming more aware of their positioning in relation to the writing. For several of the participants this sense of value originated in a "feeling of probably a kind of rough grade-band" or perhaps not even a grade-band but "a kind of top third, middle third, bottom third" (P2). She described this as a "gut feeling at the beginning", probably related to her initial impression, that had been adjusted up and down through her positioning or adjustments to her positioning. From this a letter grade "emerges". P5 sought to maintain a separation between the components during her reading and to focus on what she termed the profile of the writing in relation to the components. Her evaluation did not begin with an impression that was subsequently modified and narrowed, involving an increasing integration of cues obtained from the writing and the emergence of a grade. Rather, it consisted of an attempt to suspend her evaluation until she was in a position to consider all of the components together, trying to ensure that the influence of each on the other was limited as much as possible. She also acknowledged that if she thought that she had acquired a general impression of the writing, she tried to compartmentalise this as a separate component that she could then incorporate into her

evaluation along with other components. P4 thought that his impression of the work related to the trend or 'trajectory' that was described under positioning processes, and that this impression eventually reached a state of stability that became increasingly less likely to be altered by new information. This increasingly stable impression did not then become the grade, but served to roughly position the work before modifying it on the basis of other components of value perceived in the work. The development of a good impression did not mean that a poorer section, component or aspect of the writing would be disregarded. Rather, during the evaluation process these would be used to modify the grade representative of the impression. This suggested that P4 was able to compartmentalise aspects of the writing in order to consider these components separately when evaluating the work. However, he commented that the effect on his evaluation of an aspect or component of the writing might be influenced by how often it occurred and how salient it was, and by where in the essay it occurred.

The data suggested that participants might initially conceptualise the value of the work they are assessing in terms of a range. For example, P4 said "let's just say ... you gave it a B+, but you were ... thinking at the time that it possibly could have been an A-", suggesting that the work might have a value somewhere within that range of grades. The process of evaluation then involved narrowing the range to arrive at a specific sub-grade or percentage, and it may be different depending on whether the initial range fell within a grade-band or across the borderline. In the former case, P2 said "it's clearly in there, and it's either top, middle or bottom, in which case it gets ... a plus [e.g. C+], straight grade [C] or minus [C-]". P4 described how he thought his evaluation might be easier if there were more sub-grades between 68% and 80% and fewer sub-grades in the D grade-band. My interpretation of these data was that the process of narrowing may be less complex when the writing falls within an identifiable grade-band (rather than across two bands), and when there are sufficient sub-grades with sufficiently different identities available to the assessor. As was discussed in Part 1, the more meaningful distinctions made by assessors are between the grade-bands, the sub-grades representing variations of quality within those bands that nevertheless retain the same overall identity. This feature of the grade-bands thus simplifies the evaluation in cases where the writing falls clearly within a particular grade-band.

In cases where the assessor's impression or initial range of value for the writing straddled two adjacent grade-bands, the process of narrowing the range became one of deciding on which side of the borderline the work should be placed. This involved determining whether a piece of writing met a particular threshold requirement to move across a borderline, rather than whether it fitted a particular grade-band. It required testing identified content elements in relation to criteria, weighting of the relative importance of these, and offsetting them against each other utilising testing/deciding processes. This was in contrast to a looser determination by comparison and matching of whether the content elements in the writing were contained within the set of characteristics thought to represent a particular grade-band. Although the range of value is narrower across a borderline, it contains an either/or choice between two possibilities rather than a determination of where within a grade-band the centre of gravity of the value of the work might fall.

In performing their evaluations, the participants drew on a number of sources of reference against which they matched or compared the writing in order to arrive at a sense of its value. Some of these included criteria described in Part 1. P2 described how she compared her impression of the writing to the assignment brief, to what the student had indicated would be addressed in the essay, and to her own expectations based on her prior experience and how she had prepared to carry out the particular assessment. P3 explained that with good examples of writing she felt that she was comparing the work to her own framework of worth, judging the validity of points made in the writing with reference to her own knowledge of the subject. With poorer writing, she suggested that the value of the work was determined by more quantitative means where she reviewed her tally of instances of anything that might be considered worthy. She also noted that she used other assignments in the batch she was marking as a source of comparison. P5 tried to focus more specifically on the assignment and assessment criteria when performing her evaluation, as did P6, but also said that at times she used these more loosely as a guide to assigning values to content elements when coming to a decision. Therefore the participants appeared to draw on referents for evaluation that originated from the writing task and the criteria that related to this, from the expressed

intentions of the writers, from the writing in other essays in the batch being marked, and from their own perceptions, ideas and constructions relating to the specific assignment and writing in general.

Integrating processes

These processes refer to how the participants attempted to draw together the various components of the writing to arrive at a final judgement of its grade or mark. Some of this work is carried out by means of evaluation processes. When there are no other components identified during the course of the assessment that need to be taken into account when determining the final judgement, then the evaluation process may be sufficient. However it is possible that impressions arising out of preparation and positioning, or tallying during reading, or components resulting from counting or testing decisions may need to be incorporated or taken into account in the final judgement of the writing. The extent to which this is necessary will be a function of the salience, importance or weighting of these components, which will in turn be a result of how they are recognised or processed. It may also be a consequence of the writing containing aspects of varying quality that need to be accommodated, and may help to explain why such work tends to take longer to assess.

When talking about arriving at the final grade or mark for the writing, several of the participants made reference to the marking guidelines or rubric, often in the context of needing to complete a feedback sheet on which various aspects of the assignment were rated. For most of them the guidelines did not make a major contribution to their assessment and they appeared to make greater use of the results of the various processes described above. Several participants described how they used the comments and annotations they made while assessing a piece of work to remind themselves of features they had identified or synthesised in the work, and to help them to draw their thoughts together. P2 described how she used these comments to "guide the ticks" or how she rated items on the feedback sheet, suggesting that she did not use the guidelines and marking criteria as a part of her final integration and judgement. With respect to arriving at the final grade P7 said that "the pro's and cons are balanced up ... not scientifically", and implied that this involved "pretty subjective weighting - ok I have

got the marking criteria and so on, but at the end of the day I would say the marking is subjective because there are weaknesses and strengths in each one". This suggested that integration of the pros and cons, or weaknesses and strengths, was conducted through an internal process rather than by means of systematically relating aspects of the writing to the assessment criteria. It also suggested that assessors might need some sort of processes or mechanisms to achieve this. The exception to this subjective approach for this participant was when making a borderline decision. In this situation she said that the use of marking criteria "helps with the marking as well, because if someone is a borderline I can look at that grid and sort of calculate them, an actual mark rather than a subjective mark", and this may be part of her testing processes. This suggested that when an either/or decision is required it might be approached more mechanically by determining a preponderance with reference to external criteria.

The two participants who focussed their assessment more specifically around the assignment and marking criteria also appeared to carry out their integration in a more mechanical way. P6 felt that he constructed a grade fairly quickly by scanning and identifying content elements that fitted the criteria. Rather than including a number of components derived from some of the other processes described above (P6 did not describe any parts of his assessment in a way that could be interpreted as positioning or testing, and he appeared to simplify his evaluation process), he appeared to grade each of the typically three or four assignment criteria or task requirements, and averaged these to arrive at a final grade. When a task requirement was absent in an essay, the effect of this was only to reduce the average grade. P6 did not view the omission in terms of it potentially affecting the coherence of the entire piece of writing. P5 made use of the marking criteria in a more explicit and integrative fashion. She described herself using the feedback sheet to guide her decision-making: "I would go through writing up the feedback sheets in relation to all the specific criteria, and to guide my writing up I would be flicking through looking at the comments that I had written on the actual script and then ... I would arrive at a final decision about the mark, which is sometimes easier than others depending on ... whether it's a mixed profile or not". She emphasised that she used the profile across the criteria to guide her final judgement. She did not "sit there with a calculator and kind of work out an average ... but ... I would

use [it] to see what that profile is and to gain an overall impression". She described using the profile as "so if you have got a kind of mid-line and points below and points above so are there more points above or are there more points below". It appeared that she tried to relate aspects or components of the writing on which she had commented (and tried to keep separate in her mind until this point) to specific criteria, in order to generate a profile which she then used to gauge her final judgement of the work – the profile was for her the means of capturing the totality of the work. She also felt that this final judgement was more difficult when the profile was mixed, which I interpreted to suggest that the profile was less revealing of the final grade and required some additional processing from the assessor: "but within that certain elements may be weighted slightly heavier than others" involving both the criteria and her personal beliefs – a "bit of both depending on the piece of work".

P1 provided a more textured account than any other participant of what he thought he did when integrating his assessment to arrive at a final judgement. There also appeared to be aspects of integration throughout his assessment rather than towards the end, such that he might act on a component of sufficient importance at the time he recognised it or perceived its relevance or significance. Thus some of his integration processes might also be considered a part of other types of processes such as evaluation and possibly reading. He thought that frequently it was difficult to achieve the final integration because writing rarely fitted the marking or grade criteria which were often very general. He said "where you've got good in parts and bad in parts and where do you do your average of the two - some things you can't average" and that he used "what I try and think is a broad judgement". This appeared to involve combining various components in a variety of ways, and included averaging, weighting, adding and subtracting, compounding (a multiplicative effect) and modifying. These processes could include both the compartmentalisation of components, such that they were treated as separate from each other (when averaging, weighting, adding and subtracting), or they could be viewed as acting on, or merging with, each other during compounding or modification.

Examples of averaging described by this participant occurred when he was tallying

content elements (or “hovering”) during reading, and when these were not too disparate and did not show a preponderance towards poor or good elements. P1 said that “the averaging process ... isn’t weighted to any great extent ... I try and maintain a little clicker in my head ... and it clicks to the right every time someone says something good and it clicks to the left every time someone misses something or irritates you or misses a reference when they should have one. And provided you are just a couple up or a couple down throughout that’s reasonable”. His description suggested that the content elements to which he was responding had something of a binary value (good or poor), and that he was summing positive and negative tallies rather than averaging a range of values. Therefore this 'averaging' might more appropriately be labelled as 'adding' and 'subtracting', and he agreed when I suggested that he might be responding to the modal frequency of clicks. However he also implied that a particular content element might warrant more than one click. He offered as an example an assertion in the writing that something open to interpretation had been 'proved', and said “instantly your clicker's gone down ... below C”. This implied more than a slight shift in the 'average', and suggested that tallies might also incorporate a qualitative dimension where their combination might take on more of the nature of an average. In cases where tallies began to show a preponderance in one or other direction (“when it goes very heavily in one ... or ... another direction”), they might compound or have a multiplicative effect on the assessor's perception of the element being tallied. This effect would be greater than what might be expected from their simple sum, and would trigger a 'count'. This 'count' would then be held over until it could be combined with other components in the final integrative judgement.

With respect to 'weighting' (which P1 also described as 'averaging'), this process appeared to be related to the saliency of respective components. Rather than weighting being assigned to certain types of components (possibly arising out of an importance accorded to certain criteria), it appeared that a component was only viewed as possessing a quality worthy of special treatment when it departed sufficiently from other aspects of the writing. It needed to cross some threshold of significance or deviation from the general quality of the work or the qualities of other components in order to be 'weighted'. P1 described himself as reluctant to weight “to any great extent”, partly

because a means for doing this was typically not included in any marking criteria. However he said “if there was one very strong paragraph showing good insight ... that would very quickly move them up from a C to a B”. In his description he did not suggest that he assigned an A- or A to the component showing 'insight' which, when combined with the prior perception of a C, would result in an average of a B. Thus he did not appear to be averaging components but to be basing his judgement predominantly on the one significant component.

He followed this by saying that he would “forgive” an earlier weak component. I interpreted this to suggest that if an identifiable component was significant enough (which may involve a testing process) it may be accorded sufficient weight to result in another component being disregarded or 'forgiven', with the grade reflecting more of the former than the latter. P7 said something similar. If they have “satisfied all the important objectives and I think that they are demonstrating that they understand, ... have got meaning, ... have the knowledge, ... that overrides syntax and whatever minor errors.” In this instance the components were not being 'averaged', as the better components were not being combined with other components to yield an average or even a weighted average, but the strong components appeared to be producing a marked shift in the overall judgement. In this case one component overrode rather than modified another, removing the latter component from consideration.

The process that has been described above does not reflect weighting as it is normally understood (the assigning of greater weight to some components rather than others, but incorporating both into the final judgement), although it is interesting that P1 referred to it as weighting in his descriptions. It may be more accurate to label this as an 'overriding' process. Comments about using weighting as a part of integration in the more commonly understood way were mentioned by P5 and P7, but they did not elaborate on this in a way that might have permitted an interpretation of the process. P4 explained how he would typically have a feel for a grade by the end of an assessment, but that “I sometimes have a mismatch between my holistic feeling and the criteria I’ve set myself”. To resolve this he said “I may have set criteria but then I sort of justify to myself by saying okay I’ve set six things I might be looking at but I haven’t said how

I'm going to necessarily weight them together". These data suggested that he used a weighting process in order to resolve a mismatch between different sources of value that he used to formulate his judgement.

In describing something that he thought he could not average, which I interpreted to mean that the identity of certain components needed to be maintained and could not be influenced by other components, P1 explained how he might give a D+ for "an intact essay [which] gives no citations at all". Although it might have "C+ content" he would give it a D+ "because you haven't quoted any of the references". He followed this by saying "D+ was generous because if you don't quote references ... you've got no evidence, you've got no essay no matter how well you've written it", but that assessors are "all reasonably forgiving". This suggested that this participant would give credit for a component even when invalidated by another component. This provided evidence for the compartmentalising of components in this assessor's mind and a disregarding of the whole in this instance in order to permit the "C+ content" to be counted or accumulated additively towards the final grade. These data provide evidence of an adding and subtracting process where two components are viewed as retaining their respective identities (in contrast to the weighting process described above), and the one is offset against the other in order to arrive at a grade. As with the example of weighting above, P1 did not frame his description of the poorer component to suggest that he would arrive at the D+ grade by means of a process of averaging (by assigning a fail value or grade to the lack of references, for example). However he continued to mention the "C+ content" as a counterbalance to the absent references, suggesting that some form of averaging involving qualitative values may have been a part of the process. It is possible that a fail grade for the lack of references may have been implicit in P1's thinking about the overall D+ grade.

What is interesting about the two examples described above is that they both appear to be showing a similar process. In the former case a significant component improved the assessment by a full grade, the participant describing himself as overlooking a poorer component. In the latter instance an arguably equally significant component (although this time having a negative valence) resulted in the assessment being reduced by a full

grade, but this time both components appear to be counted in the overall assessment. This suggests that although the result of the two processes may be similar in terms of the amount of change produced in the assessment, the means by which the assessor arrived at his assessment may be different. A distinction between the two examples that may be important may relate to the respective effects on the grade of the two influencing components. In the former case the influencing component would raise the mark, the poorer component being ignored. In the latter case the influencing component is the poorer component but the better component is retained as part of the assessment process. In both cases the participant mentions assessors being "forgiving" in relation to poorer components. An interpretation of this orientation may be that assessors try to emphasise the positive and de-emphasise or ameliorate the negative in their assessment. As was discussed previously in relation to criteria, this suggests that assessors' orientation to strong writing (or components of writing) may be different to how they approach weaker components. Alternatively it is also possible that in the second example, P1 might have used a weighting process had the negative component been sufficiently noticeable to warrant it overriding the positively evaluated content. The fact that it did not may have been related to the type of content element under consideration.

A process of modification might be reflected in a situation where the qualities of one component influence the assessor's perception of another. This is different to a situation where one component may override or be offset against another. In this case, the value accorded to one component is merged with that of another such that rather than being viewed separately while their respective values are being integrated, they are combined into a potentially more holistic view of the writing. One form of evidence for this was suggested when participants described themselves as "reserving judgement" in order to reduce the possible influence of some components on their evaluation of others, as was described in the section on positioning. P1 and P4 also described how their impressions became increasingly established by writing that was of a consistent standard, influencing how they viewed the remainder of the work. These participants did not specifically describe something that might be interpreted as a modification process, but appeared to be acknowledging the possible effect of such a process. P7 suggested that

towards the end of her assessment her compartmentalisation of components diminished and they "just sort of merge". What was important for her was that the important components were present in the writing (e.g. effort, content, understanding), permitting her to view other components (spelling, syntax, minor errors) as requiring less emphasis as part of her overall integrative judgement. These data might also be interpreted as evidence of weighting (either in the sense of some components being accorded more importance than others or some components overriding others). However she did say that "they would still be, you know, ticked off for it ... [but] it's not going to make that much difference", and described the presence of good components as "extra clicks for them, which is not scientific". In the context of her describing her integrative summing up of a piece of writing "at the end of the day" and in terms of "what's important" for her assessment, I interpreted the "extra clicks" as an indicator of qualitative value rather than part of a tally and her suggestion that it was "not scientific" as a recognition of the perceptual merging of these components, some being influenced by others.

Summary of Part 2

Part 2 has described data to support the identification of six processes that assessors may employ in the process of carrying out their assessment. In addition, examples were provided of content elements and components that appear to feed into, or be operated on, by those processes. These elements and components map on to criteria in complex and not very clearly specified ways, but they appear to serve as markers or proxies for aspects of those criteria. A feature of the processes that emerged during the course of the analysis was their inter-relationship. They appeared to be neither sequential nor isolated from each other. They also did not all appear to be necessary (or utilised) to perform an assessment task and the data suggested that different assessors may employ different processes in different proportions or sequences. These processes thus represent a range of cognitive tools that assessors may draw on while carrying out their assessment. The extent to which they do so may depend on individual differences and on the reality and their constructions of the context within which they work.

Chapter 5: Discussion

Introduction

In previous studies and discussions of the assessment of writing the point has been made repeatedly that the process is obscure and poorly understood (e.g. Huot, 2002; Lumley, 2002; Wyatt-Smith and Castleton, 2005). Studies have focussed on different aspects of the process and have accorded different levels of importance to these. Several studies have also offered evidence for variation within and across assessors with respect to the final judgements they make about pieces of writing and also how they go about making those judgements (Baume and Yorke, 2002; Baird, Greatorex and Bell, 2004; Shay, 2005; Read et al., 2005). There is some agreement that the process of assessment is complex (De Remer, 1998; Shay, 2005; Crisp, 2010a), and where research points to some certainty about the process, this is often confined only to a part of a much wider process that is suggested by the findings from other studies. As discussed in Chapter 2, those researchers who have endeavoured to take a wider and more comprehensive view of the process have provided data that enables more thorough explanations of some parts of the process than others. Alternatively the data might be interpreted in ways that suggest the process may be conceptualised differently. The findings from the present study offer data that fits with some of the findings and conclusions from previous studies, and offer supplementary or alternative interpretations of these while also providing some additional findings. These data and interpretations, some of which were obtained as a consequence of the investigative method employed, contribute to a potentially more integrated understanding of writing assessment and provide fresh insights that might permit further research to explore this understanding better.

An interesting feature of the current study was that the data suggested that assessment may involve a range of constituents, but that not all of these are employed by all assessors all of the time. This accords with observations made by Webster et al. (2000) concerning the varieties of criteria, and the ignoring or adding of criteria, that their

participants employed during their assessments. My data additionally suggested that assessors might employ different aspects of assessment for different examples of writing within the same batch, and that to some extent this variation in approach may be a function of the pieces of work being assessed. Elbow (1997) and Ecclestone (2001) made similar observations. Assessors select from a range of assessment constituents and use them in a variety of combinations to carry out their assessments. Their selections may also be influenced by the assessment context – the situation, conditions, and demands both internal and external to the assessor. These assessment constituents are interrelated. Some elements influence the application of others, either rendering these unnecessary or modifying how they are employed. Some constituents work with each other, operating together or acting as necessary earlier stages for other constituents. Assessment is therefore not an homogeneous process. It is this combination of the range of assessment constituents that are employed, coupled with the variability with which they are used, that might help to explain the complexity of assessment. It may also help to explain the difficulties inherent in studying the phenomenon, and in building coherent models that are adequate to explain the process.

An aim of the present study was to explore the more private aspects of assessment that relate to the internal thinking and cognitive processing presumed to accompany the act of assessing. The data was conceptualised as falling into two major themes. Firstly, the participants made use of criteria and proxies for criteria in their assessment that could be organised into a framework of meaning that related both to the writing they were evaluating, and to the grading system they used to formulate their judgements. Secondly, they made use of processes that described how they thought about or acted on the writing and its grading, as well as how they involved themselves and directed their evaluation as they carried out their assessment. Six types of processes were identified which included both subjective positioning and the cultivation of sensitivity to possible meanings, and more objective and mechanistic identification and measurement and the application of conditions to facilitate decision-making. These processes were located within the framework of meaning provided by the criteria and operated on these criteria.

As was commented on above, not all of the participants utilised all of the processes that

were identified, nor did they view the framework of criteria in entirely consistent ways. I interpreted this as suggesting that a feature of writing assessment in the type of higher education context from which the participants were drawn, is that assessors make use of a selection of processes (and sometimes criteria) available to them. This may help to explain some of the different and idiosyncratic ways in which assessors approach their assessment that were reported by Shay (2005) and Read et al. (2005). However a more comprehensive data set might have suggested that some processes are more universally used than others, while those less frequently used might be specific to particular assessors. Additionally, my discussions with some participants focussed more on certain aspects of assessment than others, and the elicitation of more data from each participant might have shown that they utilised a wider and more similar repertoire of processes than was indicated by my findings. Therefore, as in previous studies, my findings also have gaps in the data where my interpretations are less strongly supported. Some of these gaps may also have arisen as a consequence of my choice to focus my research on an aspect of assessment that is not directly observable and dependent on the hermeneutic sensitivity of myself and my participants. Nevertheless, the data provided sufficient agreement across participants to lend credibility to the criteria and processes that were identified, while offering sufficient variation in emphasis across participants to suggest that criteria and processes might be employed in differing combinations by different assessors.

Criteria

Types of criteria

I made a distinction in my study between criteria employed by my participants that related to the grading system (the framework of meanings or values that assessors used to summarise their assessments) and criteria that related to the writing. These criteria were largely implicit and stood in contrast to the explicit assessment or marking criteria, rubrics or guidelines that are often thought of when criteria are mentioned in the context of assessment. Wyatt-Smith and Castleton (2005) also made distinctions between criteria, the grading system, and the work being evaluated, and suggested that

assessment judgements may additionally involve decisions regarding the relationships between these. Typically, assessment criteria involve a linking of, or confusion between, the meanings attributed to the grading system and to the writing, such that the meanings of one are used to explain the meanings of the other. The findings from my study suggested that one aspect of assessment may involve reconciling the meanings of one set of criteria with those of the other. The assessor approaches the assessment task with a framework of grading values with its own logic and a set of constraints that it imposes on the assessor, as well as a framework of expectations concerning writing and the writing task being assessed. While there is a loose relationship between the two sets of criteria, the data from my participants suggested that the mapping between them is not as clear as might be suggested by the descriptions in published assessment criteria. This might help to explain some of the disjunction between assessment judgement and marking criteria that has been reported in the literature (e.g. Webster et al., 2000; Lumley, 2002; Bloxham et al., 2011). From their descriptions it was apparent that the attention of my participants alternated between a focus on what they wished or hoped to see in the writing, and the assigning of a value, mark or grade to it. I suggested that this might be viewed as representing a distinction between a process of evaluation and a judgement decision point. The latter occurs when assessors shift their attention from an aspect of what is being evaluated, to consolidating their impression of that aspect of the writing, or the entire piece, with reference to the grading framework.

In contrast to the processes which will be discussed later in this chapter, the participants in my study demonstrated some commonality in the meaning frameworks they described themselves using in their assessment. Grainger et al. (2008) reported a similar finding. In part this may reflect the fact that the participants were all drawn from related subject areas and all worked in the same school of health sciences. This supports the claims made by Shay (2004, 2005) and Jawitz (2009) that assessment reflects the academic context of the assessors, and the notion of academic socialisation discussed by Crisp and Johnson (2007). This in turn suggests that the variability in marking consistency reported by Shay (2005), Read et al. (2005) and Baume and Yorke (2002) may have more to do with the processes employed by assessors than their understandings of implicit and explicit criteria. If this is true then it might help to

explain why the criteria-focussed strategies employed by Baird et al. (2004) did not result in an improvement in marking reliability. Alternatively, a greater sensitivity to less explicit criteria relating to writing and grading may be important if assessment variability is to be reduced. The hermeneutic approach to the collection of data adopted in my study permitted my participants to talk about some of these less explicit criteria. This approach addressed one of the concerns voiced by Johnston (2004) and Broad (2000) about the consequences of positivist approaches to assessment research.

Grading

For statistical and psychological reasons, assessors approach work of varying levels of quality in different ways and with different levels of confidence. My participants commented that the majority of the work they assessed was likely to fall in the C to B grade-bands and they felt that the distinction between a C and a B grade had greater significance for their students than did other grades. This may not be unrelated to the the 2:2 and 2:1 degree classifications that correspond to these grades, and which represent a distinction viewed as having high stakes for the majority of students. Elbow (1997) made a similar observation in relation to the A-B distinction in the USA. Two consequences of most work falling within the C to B grade-bands are that assessors have more experience of marking work in these grade-bands, but the task is made more difficult by the finer distinctions that assessors need to be able to make to rank scripts within these bands. Two further aspects of awarding a mark in this range are that the demand placed on the assessor in terms of consequences is small (work at the extremes of the grade range requires greater attention to be certain that it is neither under- or over-marked), but because the work is likely to be made up of both poor and good components, there will be a greater demand in terms of reconciling these components. The heterogeneity of the quality of the writing also makes it necessary for the assessor to break the work down into its constituent components and to evaluate each of these independently. Cumming et al. (2002) described how this balancing of positive and negative elements produces some uncertainty and ambivalence in assessors.

Writing towards the ends of the grade continuum tends to exhibit homogeneity of quality with fewer elements or components that deviate from that quality. Increasingly

poor work is likely to contain fewer examples of good elements, while increasingly good work is likely to contain fewer poorer elements. Such writing permits a more global and simple assessment to be made, although assessors have less experience of assessing this type of work. In terms of ensuring the appropriate ranking of pieces of work within these regions of the grade range, fewer pieces of work to evaluate make this aspect of the assessment less complicated. Greatorex (2002) and Elander et al. (2006) reported findings that indicated that the characteristics that defined different grade-bands were strongly associated with each other, and Greatorex suggested that assessors tended to conceptualise a performance more broadly rather than in terms of specific characteristics. This description fits with the approach of the participants in my study towards pieces of writing at either ends of the grade continuum, but may be less representative of how criteria, whether implicit or explicit, might need to be examined separately when evaluating work in the middle of the grade continuum. The study by Greatorex focussed on the A, B and E grade-bands at both ends of the grade range, suggesting that the level of the work being evaluated may be an important consideration when examining how assessors develop or utilise criteria. The requirements and demands placed on assessors assessing work at different points along the quality continuum are likely to be different, which suggests that there may also be differences in how they approach the task.

A further feature of the system of grading employed by the participants in my study related to how points along the grading scale were conceptualised. The grade-bands (A, B, C, D, F) were perceived to have a categorical identity as well as delineate segments of a more continuous range of percentage marks. Furthermore, the participants in my study spoke more often in terms of these grades than percentages, and tended to view the grading scale in terms of a series of ranges. These views lend support to the possibility of assessors viewing writing as having an approximate value, and grades as descriptors of this value, rather than having interval or ratio properties, as was discussed by Knight (2006). Each grade band is viewed as describing qualitatively different levels of writing and conveying a specific and distinct impression about the writing to students and other assessors. This results in the perception of a discontinuity or step-change across the borderlines between grade-bands. Elbow (1997) made similar observations

about grade-bands. The sub-grades within each band (e.g. C-, C and C+ within the C grade band) were viewed as variations within the grade band rather than as distinct categories in their own right. When writing was perceived to fall within a grade band, the determination of the final sub-grade was a relatively simple matter as it would not greatly affect the meaning that was attached to the piece of work. However, when a decision has to be made between two adjacent grade-bands, the assessment task is likely to be more difficult and to take longer. It is also likely to involve different processes as the nature of the decision (the narrowing of a range within a category versus the binary choice between two categories) is different.

Writing

The participants in my study made distinctions between the aspects of the writing that they thought they were evaluating in work at different points on the grading scale. This in turn offered evidence for how their approaches to assessment might be different. They felt that increasingly poorer pieces of work displayed increasingly less content, and that they were increasingly generous in the marks they awarded. They felt that their evaluation of this type of work was relatively simple and consisted chiefly of a quantitative counting of anything that might be relevant regardless of whether that relevance had been demonstrated by the writer. They also commented on the consequent ambiguity of the writing and the obligation on the assessor to expend additional effort on making sense of the writing. An additional feature of this level of work that might contribute to the effort required to evaluate the writing, was the tendency for assessment criteria to be phrased in terms of what should be present, rather than what was absent, as was discussed by Greatorex (2001) and Greatorex et al. (2001). This orientation towards the positive, engendered by assessment criteria, demands a longer period of scrutiny by the assessor to be certain that there is nothing in the work to satisfy a criterion. It also helps to explain why an assessor might be less discriminating in awarding a student credit for something included, given the relative novelty effect of encountering something potentially relevant in the writing.

In the case of increasingly better examples of work, the content of the writing becomes more complete and relevant to the topic and purpose of the writing. It is consequently

less ambiguous and requires less effort to read and assimilate on the part of the assessor. In these cases the attention of the assessor shifts to how the content is used, or the argument that is presented and the apparent thinking behind that argument. The evaluation is qualitative, utilises the progressive descriptors described by Greatorex (1999), is more subjective and as a consequence more difficult. This difficulty is exacerbated in cases where the writing is very good. At the upper end of the grade range the examples of writing become fewer and more unique and have no direct comparators. The judgement of the assessor becomes increasingly subjective and vulnerable to idiosyncratic preferences. As was noted by Elbow (1997), there is no one correct answer to what constitutes good writing and he suggested that it might vary depending on the perspective of the assessor. Additionally, my participants described how they encountered their own limitations and insecurities in being able to determine the worth of a piece of writing, in cases where it was very good. Assessors were less able to be explicit about their criteria, as was discussed by Lea and Street (1998), and tended to become increasingly parsimonious in awarding marks at this level of writing.

In addition to differences in assessors' approaches to the application of criteria at different points along the grade continuum, there was also a difference in the aspects of the writing to which attention was directed during assessment, depending on the quality of the work. Below the midpoint C band, assessment was increasingly content-focussed and quantitatively determined. Above this band the assessor was increasingly concerned with how the content was handled in the writing (the thinking and argument) and the assessment was framed in more qualitative terms. The borderline between the C and B bands appeared to be the point at which the transition between a quantitative content-focus and a qualitative argument-focus occured, although there may have been some focus on (qualitatively) evaluating the relevance of the selection of content as the work moved up into the C band.

With respect to the distinction that has been made between content and argument within writing, the data suggested that there were two further dimensions to each of these. The content, talked about in quantitative terms, consisted of the information, research findings, and other points and issues raised in the literature that the writer assembled to

produce the written work. However, in addition to this, the participants also expressed an interest in the extent to which the content was relevant to the purpose of the writing and the argument being advanced by the writer, introducing a qualitative dimension to the evaluation of content. With respect to argument, the participants viewed this as demonstrating the writer's understanding. This understanding related to both the content of the writing (partially revealed by the writer's selection of that content), and to the writer's appreciation of the purpose of the writing task and the need to communicate an understanding of that purpose. Consequently the argument also had two dimensions and needed to demonstrate an understanding of the topic as well as the task. This need of the writer to discern what is required of her or him was discussed by Lea and Street (1998) in the context of their concept of academic literacies. Rather than students being required to replicate in their writing only the content and style of their academic discipline as an academic socialisation model might suggest, the participants in my study acknowledged this deeper requirement for the student to engage through the writing in a negotiation with the assessor about what might constitute legitimate content and style – to argue for their argument. An important point made by Lea and Street was that the surface features of the writing (which I termed content elements and content components that served as proxies or markers for implicit criteria) could be interpreted differently by assessors from different disciplines, resulting in differences in the interpretations of the meanings and understandings that underpinned them. This suggested that the validity of the surface features of writing as a representation of the meaning of the text may be overestimated. This point will be picked up again in the section on criteria and interpretation below.

Criteria and interpretation

An added dimension to the criteria that assessors hold in relation to writing relates to the means by which they discern the presence of those criteria. I discussed these in my findings as part of processes and I described them as content elements and components. These were features of the writing that served as proxies or markers for the criteria and that mapped back onto those criteria. Content elements are characteristics of writing that are quickly and easily recognised, perhaps similar to cues recognised by decision-makers that were described by Dreyfus and Dreyfus (2005). Importantly they were not

evaluated qualitatively, but their presence or absence was more likely to be tallied and utilised as an indicator that an aspect of the writing satisfied a criterion. In some respects these content elements were similar to the procedural criteria described by Beck and Jeffery (2007) and Erling and Richardson (2010) or the technical writing skills described by Lea and Street (1998). However, my findings suggested that they also serve as proxies by means of which assessors infer the presence of criteria that they actually incorporate into their assessment. The proxies are therefore neither criteria that are used directly in the ways suggested by Erling and Richardson, nor are they the more implicit criteria that are used as a part of the evaluation aspect of the assessment. Rather they serve as pointers to the criteria that are used. Content components on the other hand are evaluated more qualitatively in relation to the educational purpose of the writing (demonstrating understanding) and the communicative intent of the writer. The components are recognised more gradually and they are more likely to be markers (as opposed to proxies) for criteria. As a consequence they are more closely linked to the implicit criteria.

This suggested that there is an interpretive layer between what might be discerned in the writing (content elements and components), and the criteria that assessors actually wish to use in making their assessments (adequate content and selection of content, and argument that reflects understanding of content and of the purpose or communicative intent of the writing). A decision has to be taken by the assessor about which content elements and components should be counted, and an argument has to be inferred from the way that the content has been used in the writing. If this is true then assessment criteria that are putatively used to improve assessment reliability may misrepresent, or only indirectly represent, what is valued in the writing, and familiarising students with these is unlikely to empower them with understanding in the manner suggested by Elwood and Klenowski (2002). Research focussing on these criteria may also be misdirected in terms of what it can contribute to an understanding of assessment. This interpretive layer between assessment criteria and assessors' implicit criteria offers some explanation for how assessment criteria can be applied in variable and contested ways, as was discussed by Condon (2009). The variability arises from both the possible mismatch between explicit assessment criteria and implicit criteria, and the interpretive

link that the assessor makes between them. It may also help to explain why assessors are less inclined to use explicit assessment criteria in their assessment, as reported by Lumley (2002), Mullins and Kiley (2002) and Bloxham et al. (2011).

The implicit criteria discussed by my participants, and the markers or proxies that I identified as possible pointers to those criteria, show some similarities to the kinds of assessment criteria reported in the literature (e.g. Lea and Street, 1998; Mullins and Kiley, 2002; Condon, 2009; Kreth et al., 2010). Elander et al. (2006) described the argument within writing as the defining feature of an essay and asserted that the essay encapsulated the argument, and Kreth et al. (2010) saw writing as a way of expressing thinking and emphasised its communication function. These views were echoed by some of my participants, and implicit in these was a belief in the contribution of these aspects of writing to the academic socialisation of students as was described by Lea and Street (1998). However my findings suggested that my participants did not discern aspects of writing which satisfied implicit criteria directly from the surface presentation of the work. Rather, surface aspects were interpreted as providing evidence for those criteria, and those interpretations emerged in a variety of ways and by means of a variety of processes which will be discussed further in the section on processes below. This helps to explain the finding by Kreth et al. (2010) relating to the focus of assessors' comments on the surface features of writing. It also explains the need described by Elander et al. (2006) for criteria to involve the integration of the surface presentation of writing with more complex cognitive functioning. The distinction I have made between implicit criteria and markers or proxies offers one possible way to conceptualise the inter-relationships between criteria described in other studies, and again suggests that an analytical application of criteria might be misdirected. Criteria that can be immediately discerned in the writing may not be directly representative of what is being evaluated, while criteria that are representative of what is being evaluated are not discernable from the writing without some interpretation.

Variations in approaches to assessment

The finding in my study that there may be different approaches to assessment depending on the quality of the work, and that assessors are forced to evaluate different aspects of

the writing at different points along the grade range, support those of Pollitt and Murray (1996), Cumming et al. (2002), Read et al. (2005), Delaney (2005) and Elander et al. (2006). Read et al. suggested that assessors might identify certain criteria with poorer work and other criteria with stronger writing, with a third set of criteria being used to evaluate writing across the entire grade range. The third type of criteria related to aspects such as understanding, relevance, argument, analysis, use of evidence, and also structure and quality of writing. In terms of the findings of my study, the poorer and better grade-specific criteria might be seen as being similar to what I described as markers and proxies in my data, while the non-specific criteria might be representative of more implicit criteria. Elander et al. and Cumming et al. suggested that at the lower end of the grade range criteria were likely to involve the more superficial skills-based and generic language and structure core criteria, while at the upper end the criteria were likely to be related to the critical thinking, argument and organisation that are relevant to the construction and communication of knowledge representative of more highly developed writing. The quantitative-qualitative distinction I made in relation to the application of criteria by my participants at the lower and upper ends of the grade range respectively, which reflected the non-progressive versus progressive distinction made by Greatorex (1999), might also be made in relation to the interpretations made by Elander et al. and Cumming et al.

The variation of writing encountered by assessors at different points on the grading scale provided an explanation for a source of assessment complexity that might contribute to variations in consistency that has been suggested by various researchers (Baume and Yorke, 2002; Baird, Greatorex and Bell, 2004; Shay, 2005; Read et al., 2005). In addition to this, the sometimes idiosyncratic ways in which assessors have employed assessment criteria, that have been discussed by Webster et al. (2000) and Ecclestone (2001), may be not so much a reflection of their individual differences, but the fact that they approach (and may need to approach) writing differently depending on how good it is. The assumption that all work of whatever quality is assessed or needs to be assessed in a similar way may not be valid. The findings of my study thus offer a possible explanation for how variations between assessors may be triggered by features of the writing, as was suggested by Elbow (1997). Rezaei and Lovorn (2010) described

these features in terms of aspects of the writing, like style and grammar, or characteristics of the writer, like command of language. My findings suggested that it may be the positioning of the writing because of those characteristics, or the student's performance as suggested by Ecclestone (2001), that may determine how it needs to be approached. Attempts to improve assessment reliability by focussing on standardising practice in ways suggested by Baird et al. (2004), regardless of the quality of the work, may be limited unless variation in the quality of the writing is also taken into account.

Processes

Introduction

One of the aims of the current study was to explore the processes of judgement and decision-making, or the processes that lead up to the point when an assessor makes, or is able to make, a judgement or decision. This was in contrast to previous studies where less of a distinction has been made between this aspect of assessment and those of understanding the writing and applying criteria and a grading system. Previous research (Milanovic et al., 1996) did not view decision-making behaviours as separate from other aspects of assessment, while Cumming et al. (2002) suggested that the rapidity and concurrent nature of decision-making behaviours might make it difficult to disaggregate them from the overall process. Although Cumming et al. and Crisp (2008a) made a distinction between decision-making and other aspects of assessment, this aspect was embedded in the total process such that its nature was not examined very closely.

As a consequence of the method adopted in the current study it was possible to examine in greater detail with the participants their understandings of how they might arrive at their judgements and decisions in relation to written work. In contrast to other studies, with the exception of those reported by Crisp (2008a, 2010b, 2011), the data in my study revealed a considerable number of processes, mechanisms, mini-procedures, operations, functionings or techniques that the participants reported themselves using as they carried out their assessments. These offered a means for providing a more detailed explication of what assessors might actually do when employing the summarily

described cognitive strategies of Suto and Greatorex (2008a), and the judgement strategies of Cumming et al. (2002). It is these processes that offer insights into the thinking that precedes and presumably contributes to the judgement, thus potentially addressing the concerns advanced by Cumming et al concerning the inaccessibility of decision-making behaviours. While the decision point might be difficult to isolate for analysis, the processes that lead up to that point are more accessible and provide a means of understanding how an assessor might be able to make the decision.

Six categories of processes

The processes identified in my findings were organised into six types or categories which appeared to characterise different aspects of the thinking of assessors. These processes also appeared to interact with each other as a part of the overall process. This contrasted with the models of Crisp (2010b, 2011) which conceptualised the overall process as a series of stages. Importantly, the data from my study provided additional evidence for the separate existence of such processes, and contributed to a description of these that might permit the development of a more comprehensive theory of assessment along the lines called for by Huot (2002). The six types of processes identified in the data were: *preparation, reading, positioning, testing/deciding*, *evaluating* and *integrating*. These were in some respects similar to the model that was put forward by Crisp (2010b) involving a sequence of five stages each incorporating a number of cognitive behaviours she had identified in the 'think aloud' commentaries of assessors. Her five stages were: a Prologue, Phases 1, 2 and 3, and an Epilogue. Cumming et al. (2002) also described an assessment sequence that involved scanning the surface features of the writing, interpretation strategies and judgement strategies (arriving at a grading decision). These judgement strategies were categorised in terms of three types of focus: self-monitoring; ideas, arguments and writing organisation; and use of language.

The prologue stage of Crisp's model included preparatory reviewing of thoughts, expectations and reminders relating to the brief, and procedural aspects of selecting scripts, which was similar to the *preparation* processes I identified in my data. In contrast to this perspective, Cumming et al. (2002) referred to these as macrostrategies

that existed outside the interpretation and judgement strategies of the actual assessment. The self-monitoring judgement behaviours described by Cumming et al. bore some similarities to the *preparation* processes as well as the *positioning* processes identified in my study.

Phase 1 of Crisp's (2010b) model incorporated five of the seven categories of cognitive behaviours she had identified earlier (Crisp, 2008a). These were reading and understanding, personal responses, social perceptions, language and task realisation. Cumming et al. classed behaviours similar to these as interpretation strategies and language-focussed judgement strategies. The first three of Crisp's strategies bore some relationship to the mechanistic *reading* processes and subjective *positioning* processes I reported in my findings. The latter two (language and task realisation), rather than being processes, might be more representative of what I described as criteria.

Phase 2 of the Crisp model incorporated her evaluation category of cognitive behaviours. Evidence for these behaviours included comments reflecting positive, negative, neutral or uncertain impressions, which could be seen as similar to the *reading* processes I labelled as identifying, tallying and counting. Additionally, Crisp included behaviours like 'weighing up' and 'comparing' as part of evaluating. In the analysis of my own data I felt that there were two types of 'evaluation', which I called *evaluating* and *integrating* processes. *Evaluating* refers to making relatively uncomplicated judgements of the value of writing where the grade gradually emerges or the assessor's impression becomes increasingly stable. This typically occurs when various components of the writing are relatively well correlated and it is not necessary for an assessor to reconcile very disparate components. These *evaluating* processes also tend to occur when a piece of writing is being evaluated as a whole, rather than being broken down into components having different values. These processes may be similar to the 'configurational' judgement processes described by Crisp (2011).

Integrating processes occur when assessment becomes more complicated, and may correspond to the processes Crisp (2011) described as 'analytic'. This arises when writing contains components of varying quality that need to be reconciled, or when

assessors attempt to maintain a separation between components while reading, before drawing them all together at the end. Rather than focus on the writing as a whole, the assessors need to consider its components and find ways of aggregating, combining or integrating their impressions of these components. The *integrating* processes (such as averaging, weighting, adding/subtracting, compounding and modifying) I identified in my data provided some explanation for how this synthesis might be achieved, and offered suggestions for how assessors might carry out the 'weighing up' and 'comparing' Crisp (2010b) included as part of her Phase 2 evaluation. The discussion, reviewing, re-assessment and justification of the mark described by Crisp as part of the fifth epilogue stage of her assessment model might also be seen as *integrating* processes.

With respect to Phase 3 in the Crisp (2010b) model, which included behaviours relating to assigning marks, evidence for these behaviours was less apparent in my data. As I discussed in my findings, arriving at a grade or mark decision was perhaps less a process and more the judgement endpoint of one or more processes. Some of the data Crisp coded under 'assigning marks' (evidence of knowledge or understanding, references to the mark scheme or assessor discussions) I regarded as relating to criteria. However there were some behaviours she included under 'assigning marks', such as 'arriving at the first indication of the mark', that might correspond to processes I incorporated under *positioning*.

Just as possible processes corresponding to Crisp's Phase 3 were less apparent in my data, Crisp did not identify separate cognitive behaviours that corresponded to the *testing/deciding* processes I described in my findings. These were viewed as contributing to *evaluating* and *integrating* processes as well as *positioning* processes, and might also offer possible mechanisms by which the 'weighing up' and 'comparing', described by Crisp as part of Phase 2, could be carried out. In identifying *testing/deciding* processes I felt that it was important to draw a distinction between these more superficial mechanistic processes and the complex synthesis of factors carried out by the *integrating* processes. The *testing/deciding* processes appeared to offer analytic tools employed by assessors to test aspects of data obtained while reading, in order to feed those decisions into the more subjectively managed processes of

positioning, *evaluating* and *integrating*. Cumming et al. (2002) made a similar observation in relation to how a judgement may be incorporated into further interpretation, although they conceptualised this judgement as a less well-defined end point of a deliberation rather than in terms of the more specific *testing/deciding* processes I described in my findings. These processes enable the assessor to simplify the assessment by consolidating a number of possible considerations into a single position. By removing these additional considerations from the assessment at this decision point in the process, the assessor has less to hold in memory as part of her or his *evaluating* or *integrating*.

In addition to the differences outlined above between Crisp's (2010b) stages and my processes, my analysis departs from the model constructed by Crisp in two respects. The first is that I have suggested that while some processes may support others, the sometimes parallel or simultaneous utilisation of processes may be more representative of cognitive activity when assessing than the sequence of stages suggested by Crisp. In her discussion of the phases of her model she recognised this, noting that behaviours characteristic of one phase might also be used during other phases. She also commented on how assessors might 'loop' between phases (particularly Phases 2 and 3) during their assessment. Cumming et al. (2002) also described how some of their participants interleaved or cycled between their interpretation and judgement strategies. These suggested that distinctions between the phases may not always be clear and that the phases may not always be followed in a progressive sequence. It may be more helpful therefore to conceptualise assessors as drawing on a repertoire of possible cognitive behaviours or processes, which I have suggested will be determined by the nature and the quality of the writing being assessed, as well as the positioning of the assessor, and perhaps the writer, in relation to the assessment.

The second difference between my organisation of my findings and the model described by Crisp (2010b) is that the distribution of processes across my six categories is more even, which might help to delineate the identity of each category more clearly. It might also suggest that the method I adopted had less of an influence in favouring some types of processes over others. This may make it easier to discern the role each category

plays in assessment, thus contributing to a more systematic explication of the process and making it easier to see how they might work together. Again, not all processes are utilised during each assessment, and the choice of which to use will in part be determined by factors that arise during the assessment process.

The ways in which processes might support or feed into other processes are illustrated in Figure 4.2 in Chapter 4. These relationships reflect something of the relative importance of the processes, or their subordinate and superordinate positioning in the framework, and the dependencies between them. Five of the six process categories feed into the sixth *integrating* processes category while four feed into *evaluating* processes. Either i*ntegrating* or e*valuating* processes are necessary to arrive at a grade or mark decision, reflecting what I have termed 'complicated' and 'uncomplicated' assessment respectively. Three processes (*preparation*, *reading* and t*esting/deciding*) feed into *positioning* processes. *Reading* processes are supported by *preparation* processes and may be supported by *testing/deciding* processes (which may contribute to the recognition of a content component), while some initial *reading* processes may contribute to *preparation. Positioning* processes feed back into r*eading* processes (to re-calibrate the approach to reading in the case of poorer examples of writing). There were no data to suggest that any of the other processes provided input to the t*esting/deciding* processes. The dependencies between the categories of processes suggest that those requiring subjective synthesis as part of their functioning, particularly *positioning*, *evaluating* and *integrating*, are supported by more objective, analytical or mechanistic processes (*reading* and *testing/deciding*) that appear to be more independent and less determined by subjective processes. The relationships between the processes will be examined further in the discussion that follows.

Preparation processes

Preparation processes enable assessors to perform two main functions in relation to their assessment. The first of these involves the cultivation of openness to possible meanings in the writing, and a sensitivity to the multiple relationships between words and the thoughts they are intended to express. This reflects two of the points made by Scott (2005) and the emphasis placed by Read et al. (2004) and Contreras-McGavin and

Kezar (2007) on the interpretation task of the assessor. The openness described by my participants also referred to a belief that their expectations of the writing might evolve during the assessment, and that the writing might alter what the assessors thought they were looking for in the particular assessment. This recognises that participants see the students' writing as a representation of the students' understanding of the assessment task, as suggested by Mullins and Kiley (2002) and Elander et al. (2006), and as a way for the assessor to try to understand the students' understanding, which was similar to a finding by Kreth et al. (2005). It also suggested a willingness to recognise that the writing would reflect the context of the writer as well as that of the assessor, as discussed by Knight (2006), and the importance of a shared understanding between assessors and assessed (Baume and Yorke, 2002). Lea and Street (1998) described this orientation to writing as reflecting an academic literacies approach, where assessors are open to differences in perspective and the negotiation of the contested nature of those differences, and recognise that writing practices need to vary to accommodate them.

The second function of *preparation* processes is to cue assessors into how they should approach the assessment. This relates both to the amount of care with which they need to approach an assessment and whether they are likely to evaluate the work quantitatively or qualitatively. In both cases this is likely to be determined by the quality of the work. Poorer examples of writing require greater attention and care from assessors, and they are more likely to look for anything to which they can award a mark regardless of how well it might fit with the piece of writing as a whole. With better examples of writing less attention is required, and the evaluation is likely to be more qualitatively focussed on the whole and how well the components of the writing have been integrated into this whole.

Reading processes

The reading behaviours described under r*eading* processes were unremarkable and similar behaviours have been reported in other studies (e.g. Lumley, 2002; Read et al., 2004; Crisp, 2008a, 2011; Suto and Greatorex, 2008b). However, unlike the interpretations offered in these other studies I did not include processes supporting deeper meaning-making as part of reading. I felt that it was important to make a

distinction between the more mechanistic aspects of scanning writing and identifying content elements within it, and the integrative, holistic and qualitative processes needed to discern or construct the meanings implied by the writing.

In addition, I included as part of the r*eading* processes the tallying and counting of content elements, which my participants appeared to maintain as distinct from the use they made of these tallies and counts as part of their evaluation. The *positioning*, *evaluating* and *integrating* processes appeared to act on the results of the tallying processes rather than subsume them. Other studies have not made this distinction. Conceptualising this crude form of measurement (tallying and counting) as an input to the more subjective processes might contribute to an understanding of the sorts of values that an assessor works with when carrying out those processes. Not only might it be useful to conceptualise reading and meaning-making as distinct from the act of judgement of a piece of work, as suggested by Lumley (2002) and Mullins and Kiley (2002), but it may also be reasonable to suggest that there are objective and subjective dimensions to reading and meaning-making.

The data in my study did not provide much evidence for a separate process of meaning-making, or my participants did not discuss this aspect of their assessment in these terms. Rather they spoke in terms of meaning frameworks which they employed as part of their assessment and which I have discussed under criteria. This may be a function of how the data were collected. Previous studies that have identified meaning-making behaviours in their participants (Lumley, 2002; Delaney, 2005; Suto and Greatorex, 2008b), or that have provided data that might be interpreted as meaning-making (e.g. Crisp, 2008a), employed a 'think aloud' procedure to collect their data. Although these studies made it apparent that assessors engaged in meaning-making, the means whereby they attributed that meaning was less clear. The data relating to the frameworks of meaning or non-explicit criteria employed by the participants in my study offered some suggestions for how they might arrive at the meanings they construct of the writing. As this involves interpretation and subjectivity (Read et al., 2004; Scott, 2005; Contreras-McGavin and Kezar, 2007), I have suggested in an earlier discussion that this is more likely to take place within the *positioning, evaluating* or *integrating* processes. In

contrast to the suggestion that meaning-making is a stage of assessment that takes place before evaluation, assessors may make use of their meaning frameworks throughout the subjective processes of *positioning*, *evaluating* or *integrating*. An important implication for future research will be to explore further the ways in which the meaning frameworks relating to writing and grading are employed during these more subjective assessment processes, and the possible reciprocal ways in which these meaning frameworks are constructed.

Positioning processes

There were two features of the *positioning* processes identified in the data. The first was a continuation of the cultivation of openness discussed in relation to the *preparation* processes, which several of the participants referred to as 'reserving' or 'suspending' judgement. The second was participants' sensitivity to the trajectory of the writing or their increasing certainty regarding the worth of the work. Although these seem contradictory, I interpreted there to be a shift in emphasis from the former to the latter as the assessor progressed through the piece of work. The operation of these processes was also influenced by the quality of the work, and was in some respects linked to the results of the *preparation* processes and how the assessors felt they needed to approach the work. With better examples of writing, assessors were able to position it earlier in terms of its value, and read with the expectation that it would be good throughout. Their suspension of judgement was confined to maintaining an awareness that there might be something later in the work that might disconfirm this impression. The consequence of this *positioning* was that the assessors would be able to focus on the whole, which would support their *evaluating* processes in arriving at a mark or grade. With poorer writing, assessors needed to focus more on the components of the writing and reserve their judgement, perhaps to the end of the assessment, before being able to arrive at a sense of the value of the work. In this case the *reading* processes might be altered to favour a slower and more careful reading of the work, and tallying/counting of content elements and components, in order to incorporate these into the more complicated *integrating* processes needed to arrive at a grade.

An interesting aspect of this difference in approach to better or poorer pieces of writing

was the assumption of worth in students' writing by the participants. When writing looked good, they worked on the assumption that it would be good unless proved otherwise. When it was poor, they reserved judgement in the hope of something good emerging that might enable them to award a better grade than was initially suggested by the work. The different approaches also suggested that, where possible and justified, assessors might use a simpler and faster approach to their assessment, providing an example of how the nature of the writing being assessed might alter the processes used to assess it, in the interests of simplifying and speeding up the task for the assessor. This aspect of how processes might be used will be explored further in the next chapter.

Testing/Deciding processes

These processes appeared to arise from the need of assessors to make a choice about a particular content element or component of the writing, in order to be able to incorporate this decision into their *positioning, evaluating* or *integrating*. They reflect a binary choice concerning whether something is present or absent in the writing, meets a criterion, or has departed sufficiently from a criterion to require attention, or whether or not one element or component outweighs another. They describe the point at which an aspect of the assessment crystallises in the mind of the assessor, taking on a particular identity that becomes incorporated into other aspects of the process. The assessors are not concerned with qualitative issues such as the nature of a criterion (which may already be represented by a proxy), or the extent to which one element might outweigh another, but rather whether or not they should include something in their assessment. They are sensitive to disruption in their reading, and once this disruption or disturbance reaches a critical level, they are able to take a decision regarding the relevant element of the writing. In order to test or make a decision regarding an impression of the writing, the participants described themselves making use of both internal and external sources of reference. With respect to the former, some of these might arise from the initial orientation to the writing achieved through the *preparation* processes, and the expectations arising from the assessor's prior experience of assessment and beliefs about the purpose of writing – their implicit criteria (Bloxham et al., 2011). External sources of reference may relate more closely to the requirements in the brief for the writing, or to aspects of the explicit marking and assessment guidelines.

The *testing/deciding* processes reflect a simplification of one of the aspects of the assessment to arrive at a decision, in order to release the assessor from having to continue to monitor the information in the writing that engaged them prior to the decision being taken. This then permits a reduction in the number of factors the assessor needs to hold, consequently reducing the complexity of the assessment. These processes offer a means whereby the amount of information needing to be taken into account by the assessor might be reduced, offering one explanation for how human assessors might manage the complex process of reducing the number of discriminations and integrating the many cues that Rust (2007) suggested were likely to interfere with the reliability of assessment.

Evaluating processes

As suggested by the label, these processes relate to the means whereby assessors arrive at a sense of the value of the writing, or the value emerges or is developed by the assessor. It occurs when the assessor's positioning has become stable and tied to a grade band or a wider range on the grading scale, and is reached after adjustments have been made through *reading* and *positioning* processes and the results of *testing/deciding* processes operating on aspects of these. The assessor's relatively stable position with respect to the writing does not then map on to a grade but, through the process of *evaluating*, it is modified by other components the assessor considers important in order to arrive at the actual grade. Alternatively, where an assessor has endeavoured to avoid taking a position with respect to the writing, evaluation involves the combining of values attributed to the components. In both cases, these processes appear to occur in instances where the values attributed to the components do not differ too markedly from each other or from the assessor's position with respect to the writing. I have termed this 'uncomplicated' assessment, where the assessor's various impressions converge fairly easily towards a grade. Having identified the broader grade range within which the writing is perceived to fall, the assessor then considers the various components to narrow this range and arrive at a final grade or mark. Bloxham et al. (2011) described a similar strategy. In terms of the discussion by Knight (2006) of the observer-relative and non-determinate nature of assessment judgements, the width of the initial grade

range might be seen as representing the extent of the non-determinacy, which then becomes less non-determinate as the *evaluating* processes permit some narrowing of the grade range. When the range falls across a borderline the task of the assessor is to decide which side of the borderline the work should fall. In this instance the assessor may make more explicit use of *testing/deciding* processes to support this decision, or if this is insufficient to permit a decision to be made (the decision is more 'complicated'), the assessor will need to make use of *integrating* processes. Thus *evaluating* processes involve two kinds of decision, whether the writing on balance fits a particular grade category or whether or not it meets a certain threshold requirement to determine on which side of a borderline it should fall.

Integrating processes

In contrast to the *evaluating* processes, the *integrating* processes are more likely to be used when the content elements and components, and the impressions and values arising out of these, do not converge on a grade range but consist of less easily reconcilable constituents. In this case, the assessor requires a means for capturing the totality of the work or for making sense of the profile of variation across the piece of writing. Essentially this requires ways of combining the components, which can include both compartmentalising components in order to add, subtract, weight or average them, and allowing them to influence each other, such that they act on or merge with each other (which I termed 'compounding' and 'modifying'), or one overrides the other.

An interesting aspect of my participants' descriptions of *integrating* processes was that the meanings of the words they used to describe what they did, were not those typically attributed to those words in other contexts. This discrepancy may contribute something to an understanding of how assessors conceptualise their thinking, by offering an insight into the mental frameworks that they employ. Thus 'averaging' appears to describe a situation where both relatively poor and good instances of the less important content elements occur with approximately equal frequencies and an even distribution (i.e. no clusters of one or the other type). The impression derived from this process is positioned roughly midway between the points defined as good and poor, which describe the two states in a binary system, and reflects a process of summing the two

states or adding and subtracting counts from a tally. The good and poor elements are not assigned values that vary on a continuum, and the summed value is not modified by being divided by the number of instances tallied. Importantly this appears to occur only when the two states are approximately equal in number and evenly distributed. When they are not, and the tallies begin to show a preponderance in one or other direction, they appear to have a multiplicative or compounding effect on each other such that their importance becomes greater than might be anticipated from just their sum. This then triggers what I termed a 'count' (as opposed to a tally), signifying a component of greater importance that has something of an independent identity. This component is then held over in order that it can be considered in conjunction with other components, or the impressions and values attributed to those components.

The term 'weighting' was used by participants when referring to a process which involves recognition of the perceived importance of certain components. A component is weighted because it differs from other components. It needs to cross some threshold of significance, or deviate from the general quality of the work or the qualities of other components, to be treated in this way. Determining this might require a *testing/deciding* process. The effect of a component being weighted in this way (something particularly good, for example) can be that another (poor) component might be disregarded or 'forgiven', the final grade reflecting more a judgement of the former than the latter component. This is in contrast to the more conventional meaning of weighting whereby both components would be incorporated into the final grade, but in differing proportions. This process appears to reflect a shift in the assessor's judgement, such that one component overrides or removes from consideration another component. Consequently it may be useful to refer to this as an 'overriding process' in order to distinguish it from the more conventional definition of the term 'weighting'.

It appeared therefore that some assessors adopted processes that involved privileging some components over others, or justifying the removal of some components from consideration in the final assessment process, rather than making use of more mathematically applied processes such as averaging and weighting. This suggested that *integrating* (in the sense of arriving at a judgement that incorporates a number of

disparate components) may involve processes that crystallise the identity and value of components, and determine whether these components should be retained or discarded in the final judgement. If true, this would suggest that in order to carry out this more complicated judgement, assessors might manage the process by simplifying or reducing the value dimensions of components, as well as reducing the number of components that need to be considered. It also implies that assessors may not be capable of integrating the complex array of impressions acquired while reading a piece of writing, in a manner approximating the mathematical objectivity implied by the use of terms like 'averaging' and 'weighting'. This may be an example of the misuse of terms employed to describe some of the processes of assessment, in the same way as was suggested by Knight (2006) in relation to the meanings attributed to the numbers generated by the processes, such that an impression is conveyed that the assessment is more scientific and objective than is actually the case. Similarly this interpretation supports the contention by Rust (2007) that human assessors lack the capability to systematically integrate the number of cues involved in a typical assessment task. Given that such simplifications introduced by assessors represent an approximation of the qualities of the writing, and that these simplifications may not always be conducted in the same way by different assessors, the existence of these processes is likely to increase the difficulty of reaching agreement between assessors that was discussed by Rust.

The data supporting the interpretations of the *integrating* processes was obtained largely from two of my participants. They are therefore limited, and a more directed investigation of this aspect of assessment would be an important implication for future research. However, data obtained from several of my participants supported the interpretation that some assessments are more complicated than others, and that this tends to be in situations where the assessor has to manage a set of impressions that do not agree easily with each other. The findings from my study offer some suggestions for how this might be carried out.

Theories of decision-making

In the context of the findings, consideration needs to be given to the applicability of the type of theories of decision-making advanced by Tversky and Kahneman (1974), Klein

(1997), Sloman (2002), Gilovich and Griffin (2002), Laming (2004), Evans (2007) and others, to the sorts of decision-making that takes place in assessment. Suto and Greatorex (2008b) offered the argument that it might be possible to regard some of the thinking of assessors as rapid intuitive System 1 behaviours, while other thinking was more representative of slower reasoned System 2 behaviours. One possible reason for the attractiveness of this type of theory for explaining decision-making in assessment may be the difficulty that has been encountered in describing what it is that assessors do when assessing, and in gaining access to their tacit understanding of the process. As participants have found it difficult to explain what it is they do, or provide little data about their decision-making when thinking aloud, it is tempting to describe their thinking as intuitive. With regard to the aspects of their thinking that they are able to describe, they have done so in such a way that their descriptions might be construed as evidence of reasoning behaviour.

Suto and Greatorex (2008b) identified two System 1 and two System 2 cognitive marking strategies that were used by their participants. The data from my study and the studies by Crisp (2008a, 2010b) and Cumming et al. (2002) suggested that assessors employ many more processes or strategies than these. More importantly, the interpretations that I have made of my data, and their relationship to the model described by Crisp, suggested that the strategies interpreted as rapid and intuitive System 1 behaviours by Suto and Greatorex ('matching' and 'scanning'), might more accurately be viewed as relatively superficial and simple mechanical *reading* processes involving the rapid identification of surface features of the writing. That is, they are not decision-making behaviours at all, but rather a means for registering or accumulating information on which decisions might be based. Alternatively, scanning and matching may represent the process whereby the markers and proxies that are indicative of criteria are identified. These proxies are not the contextually-triggered and unconsciously identified substitutes for less accessible complex concepts or evaluations that are described in the literature on decision-making heuristics. Rather, they are features of the writing that are consciously recognised by assessors as being representative of qualities they value. The data also offered suggestions for how this information might be held or summarised by means of tallying or counting, further

confirming its identity as something capable of additional processing.

The participants in my study did not appear to utilise attribute substitution to simplify their processing of multiple and complexly interacting instances of information, by employing heuristics. Rather, they appeared to make use of more deliberate alternative processes such as tallying and counting as part of *reading* processes, and the application of conditions or criteria as part of *testing/deciding* processes, in order to aggregate or 'chunk' information, or eliminate information from the matrix of data being processed. The subjective categories of processes also provided some evidence of simplification. For example, in *evaluating* processes, the participants appeared to identify writing as falling within a certain range on the grading scale, before then focussing on narrowing that range to arrive at a specific grade. In the case of *integrating* processes they used mechanisms to aggregate or combine components that had been discerned in the writing. The processes I described as *preparation, positioning, evaluating* and *integrating* bore some of the characteristics of System 2 behaviours, suggesting that writing assessment may involve at least an attempt at deliberate analysis and synthesis and quasi-mathematical processing, even though assessors may struggle with the complexity of the task. However, unlike the serial nature of System 2 behaviours described in the literature (e.g. Evans, 2003), a feature of my data was the considerable amount of information my participants felt they had to process, and their tendency to select from the processes available to them those that appeared to fit the assessment demands of the particular piece of writing. Therefore it may be possible to argue that the assessment of writing, as discussed by my participants, may reflect neither System 1 nor System 2 decision-making.

A feature of the *preparation* and *positioning* processes identified in my participants was their deliberate cultivation of openness to what the writer might offer, rather than attempt to find within it something representative of something previously encountered, as might be suggested by the the Recognition-Primed Decision model described by Klein (1997). Similarly, the processes categorised as *evaluating* and *integrating* did not display the sort of unconscious expert judgement suggested by Dreyfus and Dreyfus (2005), or belief-based judgement discussed by Evans (2007). My participants

described themselves as anxious to eliminate from their assessment anything that they were unable to rationalise, or that might reflect bias or undue influence arising out of their emotional response to the writing. The processes identified in my data indicated that the participants placed an emphasis on engaging with the writing, discerning its meaning and attempting to consciously integrate a variety of values attributed to the writing, in order to reach their judgement of the grade it should be awarded. They did not appear to substitute more easily accessible surrogate values, in the manner suggested by heuristic decision-making theorists (Stanovich et al., 2008).

This discussion suggests that it may be important to make a distinction between judgements and decision-making that have as their endpoint the determination of a value for something, and those that involve the making of a choice, often for the purpose of determining a course of action. There may also be a distinction to be made between decision-making which requires an assessor to manage a degree of uncertainty regarding the values of aspects of writing, and how these values ought to be combined or integrated, and that which involves the assigning of probabilities in the context of uncertainty to the occurrence of certain events or variables.

Chapter 6: Conclusion

Introduction

Previous research has suggested that the process whereby assessors arrive at their determination of a mark or grade for a piece of writing is complex and unclear, and has emphasised the importance of developing a theoretical understanding of the nature of this process. The findings from my study have provided an additional source of data to contribute to this understanding, and I have discussed how this data might be organised to provide a perspective on the complexity of the assessment process. A feature of this organisation has been a distinction between criteria and processes, and a consideration of how processes might operate on, with, and within, the frameworks of meaning constituted by the criteria. While previous work has given consideration to both of these aspects of assessment, the distinctions made between them in descriptions of assessment strategies have been less clear. With respect to processes, I have also made a distinction between objective and analytical processes (*reading* and *testing/deciding*) that relate to the surface features of the writing, and the subjective processes of *preparation, positioning, evaluating* and *integrating* that require greater synthesis. Within the constraints of the limited extent of the study, the data have provided an expanded description of these constituents of assessment and their relationships with each other. This contributes to an understanding of the process of assessment and provides pointers for future research.

Constituents of assessment

A pervasive aspect of my findings, and a contributing factor to the complexity of the assessment process, is the diversity of the constituents of assessment that are used by assessors. This diversity arises from: the ways in which various forms of criteria have come to be conceptualised; the experiences and socialisation of assessors; and some of the peculiarities of the assessment task itself. To a certain extent assessors are able to

exercise choices in how they conceptualise criteria or select or apply processes, but they are also constrained by features of the writing they assess such that their choices are restricted, or they are forced into certain modes of assessment.

As was demonstrated in Chapter 4, the assessor is likely to encounter greater homogeneity in the constituents of the writing at either end of the grading range, whereas towards the middle there is likely to be a mix of poorer and better constituents that will require greater effort to reconcile. As a consequence, an assessor's task is simpler when the writing is good or poor, and because of its homogeneity it is amenable to being considered as a whole. When it is neither good nor poor but made up of a mix of these constituents, the assessor is forced to increase the complexity of the assessment process by identifying the constituents and then utilising processes that will permit her or him to combine or integrate them in some way. In addition to this, the data suggest that the assessment of poorer pieces of work is likely to involve the quantitative handling of content aspects of the writing that are more likely to be discerned in the surface presentation of that writing. The assessment of better examples of writing is likely to consist of a qualitative appraisal of the thinking and argument presented through the writing that will require a greater degree of interpretation from the assessor. An implication of this is that the way in which the assessment is performed will, in part, be determined by the demands of a particular piece of writing, and therefore that not all pieces of work in a batch will be assessed in the same way.

The effect of the quality of the work on the criteria and processes employed in its assessment suggests that it is insufficient to conceptualise writing quality in terms of criteria qualified by progressive descriptors. Each piece of work may demand a selection of criteria and processes specific to its own unique presentation. At the very least it may be helpful when drawing up assessment criteria to identify separate (although possibly overlapping) sets of criteria for work that is broadly poor, middling or good, or to recognise that some criteria are more or less applicable depending on the quality of the work. This might assist novice assessors as well as make the process more transparent for students. In terms of research involving criteria, any findings in relation to work of a particular quality may not be applicable to writing that is better or

worse than that. This may help to explain some of the inconsistencies in the application of criteria reported in previous studies. The implicit criterion frameworks described by the participants in my study demonstrated similarities. Therefore it is possible that the application by assessors of different criteria in different ways may be a function of the characteristics or quality of the writing the assessors have been given to assess, rather than variations in the assessors' conceptualisations or applications of the criteria.

Simplification in assessment

Previous research has identified variation in the judgements made of writing across assessors, and across repeated assessments by the same assessor. Reasons offered for this have been the complexity of the process and the multiplicity of variables involved. Studies aimed at improving the reliability of assessment have typically sought to find ways of simplifying the process or the types of variables involved, although some critics have argued that this alters the process in a way that compromises its validity. However, an interesting thread that runs through the findings of the current study is that much of what assessors do might be interpreted as simplification, or pursuing the simplest course, and this is what enables them to manage the complexity of the task. Variation in the complexity of assessment arises from the effect of the writing on both the framework of criteria, and the processes of assessment carried out by the assessor. Typically the assessor will follow the simplest route to arrive at an assessment unless forced to do otherwise, and may structure criteria, or employ processes in such a way that they serve to reduce the variables that need to be considered, to simplify the process.

An aspect of simplification evident in the participants' constructions of criteria is their tendency to conceptualise the grade range in terms of distinct categories of grade-bands rather than a continuum. This reduces the number of discriminations the assessor needs to make. As the overall worth of the work has been determined, the choice of sub-grade within the band becomes a low-stakes and relatively simple decision. Where the initially-determined range of marks falls across the borderline between grade categories

and the work contains constituents from both grades, it becomes necessary for the assessor to increase the complexity of his or her assessment. This involves both a need to separately identify those constituents, and to determine whether sufficient of these exist to permit the mark or grade for the work to cross the borderline. Thus, in terms of the use that is made of the grading categories, the assessor pursues the simplest course unless the nature of the writing and the demands of the task necessitate a more complex approach.

The assessor's *positioning* in relation to the writing and its assessment is also influenced by the writing, which in turn will determine how he or she approaches the task. Again, it is the quality of the writing that cues the assessor into how to approach it. Better work allows earlier positioning, permitting a more global reading of the work while remaining alert for anything that might disconfirm the favourable impression. The assessment task is consequently relatively simple. In contrast to this, a piece of work that is heterogeneous with respect to quality requires the assessor to identify and hold components of the writing of different quality. This results in a more complex and slower assessment process, where inevitably the consolidation of the assessor's position occurs later in the reading of the work. Not only do different types of writing demand the application of different criteria, but the approach, speed and complexity of the assessment is also affected. Again, the assessor pursues a simpler approach to assessment unless forced to act otherwise by the quality of the work.

Reading and *testing/deciding* processes permit an assessor to consolidate a number of content elements, or to crystallise an aspect of the writing into a value or decision, that can then be incorporated into other aspects of the assessment process. This reduces the number of assessment constituents that assessors need to hold and incorporate into their *evaluating* or *integrating*. Therefore they are able to give greater attention to other features of the writing. The *reading* and *testing/deciding* processes appear to be processes employed deliberately by assessors, under their conscious control, in order to simplify their assessment, thereby helping to reduce the cognitive demand of the assessment task. In addition to the assessor making use of the least number and least complicated of processes to simplify the assessment process, they also have recourse to

mechanisms that can introduce simplification. It might be possible therefore to speculate that one principle governing the way assessors approach their task is that they will seek to simplify the process wherever possible. This includes choosing or framing criteria in ways which simplify them, only using more complicated processes when less complicated ones are not possible, and having *testing/reading* and *deciding* processes as mechanisms for simplifying the assessment task.

Commentary on previous work has suggested that restrictions on the reliability of assessment may be a consequence of the amount of information that has to be processed, and the complexity inherent in the task. My research suggests it may be possible to infer that complexity resulting in a reduction of reliability in assessment may reflect a presence in the writing of elements or components that do not easily permit aggregation by means of *reading* or *testing/deciding* processes. Alternatively complexity may exist because of a lack of a mechanism, or a lack of experience or inclination in an assessor, to permit simplification to take place, or a constraint on the assessment process (perhaps by the requirements of a formal assessment protocol or guidelines) that prevents simplification. Taking as a starting point the perspective that complexity may in part reflect the absence of a mechanism for simplification, an implication for future research would be that the investigation of simplification mechanisms might reveal something of the nature of the complexity of writing assessment.

Use of criteria

In addition to there being a distinction in my data between explicit assessment criteria and the more implicit criteria that assessors may use as well as, or in place of, explicit criteria, my findings have suggested two aspects of criteria that have implications for how they might be used in assessment. First, assessors may hold implicit criteria with respect to both the writing and the grading system – both are represented in the mind of the assessor. Maintaining this distinction, and a sensitivity to how these two sets of criteria interact with each other, may be important for a fuller understanding of how the

link between the characteristics of the writing and the grade are made. The second aspect relates to the extent to which evidence for the criteria may be discerned directly in the writing. My study suggests that while assessors recognise and identify certain features in the writing, these serve as proxies and markers for the criteria rather than the criteria themselves. This distinction between a signifier and a signified may contribute to a more complete conceptualisation of the link between the writing and the sense that is made of it. In the case of proxies, the interpretive link to the criterion may have arisen as a consequence of a correlation, while markers may be the result of a more extended interpretive synthesis. Consequently, features discerned in writing may not be directly representative of what is being assessed, and features that are closer to being representative are not discernable without interpretation. In both cases there is an interpretive layer between the writing and the criterion that is likely to be contested between assessors, and is a further source of variability between assessors. This implies that it may be important for future research to acknowledge that what is recognised or identified in writing may not be directly related to criteria, regardless of whether those criteria are explicit or implicit. In addition to the distinction that might need to be made between explicit assessment criteria, implicit criteria and the proxies or markers that represent criteria (and which may look like, or even be, explicit criteria), consideration also needs to be given to the interpretive links between them. Without this consideration, tight specification of explicit assessment criteria and rigorous training of assessors may have a limited effect on improving assessment reliability.

Processes

My interpretations of my data suggest that varieties of information processing take place before an assessor can make an assessment decision, and there are a number of processes that are employed as the assessor approaches a decision point or judgement. These processes exist in superordinate and subordinate relationships to each other, and they are interdependent and interact with each other. The processes characterise different types or modes of thinking or orientation while assessing, the more subordinate analytic processes supporting the synthesis carried out by means of the

superordinate subjective processes. The selection of processes used by the assessor also varies, depending on the complexity of the assessment and the qualities of the work being assessed.

Rather than being carried out sequentially, the demands of the assessment task determine which processes are activated, and how they are employed and combined, as the assessor moves through the assessment process. The demands include the nature and quality of the writing, its transparency and the ease with which markers or proxies are identified, the relationship between markers and proxies and implicit criteria, where on the grading scale the writing appears to fall, whether it fits one or more than one grade category, and whether it can be approached in a complicated or uncomplicated way. These demands arise in no particular sequence and may to some extent depend on the constructions and experience of the assessor. Thus writing assessment may be driven to some extent by the writing and the uniqueness of how it presents to the reader, and how the assessment itself evolves. If this is true, then it may be difficult to formulate a basis for standardising this type of assessment. It may be more productive therefore to examine how the variable application of criteria and processes might be incorporated into assessment rather than try to reduce that variability, perhaps by eliminating some of the constituents of the process that are critical to its credibility.

Although there were some similarities in my findings to previous research, I organised processes, and conceptualised their relationship to criteria, differently to the organisation suggested by earlier work. I disaggregated some of the strategies identified in earlier studies into several categories of processes. The analysis of my data also suggested that strategies might be reconceptualised as consisting of criteria and processes. This might permit distinctions to be made between aspects of strategies that relate to frameworks of meaning (or criteria), and those that are more likely to describe how assessors carry out their assessment. In particular, my approach to the collection and analysis of the data suggested that while meaning-making is a part of assessment, this may not be a distinct process carried out prior to evaluation or judgement, but rather describes the application of criteria while engaged in processes such as *positioning, evaluating* or *integrating.* If there is some validity to this interpretation, an implication

for future research will be to explore in greater detail how meanings and meaning frameworks operate within assessment processes.

I was able to discern more categories of processes in my data than had been identified in previous studies. I was also able to identify a number of processes that characterised each category, providing a distinct identity for each of the categories. Thus it was possible to delineate my categories more comprehensively than was the case for some of the assessment strategies identified in earlier studies. These expanded descriptions of processes offer a more complete explanation of aspects of assessment previously considered to be tacit. As was discussed previously, the credibility of these categories and the processes that define them rests on the credibility of my interpretations of the data. These interpretations were limited by the extent to which I was able to confirm some of the processes across several of the participants. In part this was a consequence of practical limitations on the amount of data that it was feasible to collect, but it may also have been because of difficulties in bringing tacit aspects of assessor behaviour to conscious awareness. Nevertheless, the categories of processes that I arrived at provide another possible framework for examining data collected in the future to explore the validity of my theoretical explanation for the assessment process.

The distinction I made between *evaluating* and *integrating* processes raises the question of whether the value of a piece of writing emerges from a reading of it, or assessors arrive at or develop a sense of the value of the writing. Does the writing reveal its value to the assessor or does the assessor need to act on the information presented by the writing to arrive at its value? An implication of my interpretation is that assessors may do either, depending on the nature of the work. When the quality of the work is relatively homogeneous its value is more easily signalled by the writing without requiring too much effort on the part of the assessor. The evaluation arises from the global impression imparted by the writing, and the task for the assessor is either the relatively uncomplicated one of narrowing the range within which the writing falls to a specific grade or mark, or the slightly more taxing one of deciding on which side of a grade boundary it falls. When the work is not homogeneous, assessors are faced with the more complicated task of integrating the disparate components that make up the

writing in a much more active way, and they make use of more complex quasi-mathematical processes to do this.

The findings of my study provided rudimentary descriptions of some of the possible quasi-mathematical processes employed by assessors, and how these might be conceptualised by some of them. An important implication for future research is that there is a need for a more systematic investigation of this aspect of the assessment process, and how components that are difficult to reconcile with each other are integrated. The discrepancy between what the participants appeared to be doing when integrating, and the terms they used to describe these processes, suggests that a more accurate description of the processes might begin to provide a more authentic and testable account of what it is that assessors actually do. This then might contribute to a better understanding of how and why some components of writing are privileged over others, some are discarded while others are retained, and some are merged with each other in apparently complex ways. The effect of these processes is a gradual reduction in components, in a manner more complicated than that performed by *testing/deciding* processes. It might be argued that the *testing/deciding* processes involve a reduction in the number of value dimensions in the components of the writing, while the *integrating* processes then reduce the number of components. Therefore there is a series of successive approximations performed by assessors as they move through their assessment, rather than an all-at-once integration of many cues or components.

If this conceptualisation of assessment as consisting of a progressive reduction in the constituents of the process is true, then it is likely that once an element or component of the writing has been eliminated or absorbed into another component, the probability of it being reintroduced is small. This may be one mechanism by which biases may be introduced into the assessment process, and may help to explain some of the inconsistencies or illogicalities in the ways that assessors regard some aspects of writing in comparison to others. An example of this was some tendency reported by the participants for them to ignore or eliminate poor components from consideration when a very good component was present, but to retain a good component to modify a very poor component. Their overall orientation towards the writing becomes fixed to a

certain extent. It is possible that this arises from an earlier decision point or simplification which then becomes difficult to alter later on in the assessment. This may be partly because once the simplification has taken place, the data that might contradict the orientation are no longer available to the assessor.

Professional implications

The findings of the study suggest three professional implications for assessors working in the kind of higher education context I have described in this study. The first two have direct practical implications for an understanding of the assessment process that might be used by assessors as part of their practice. They contribute frames of reference that might help assessors to communicate with each other their understandings of assessment, and provide suggestions that inexperienced assessors can draw on to develop their practice.

Firstly, the findings of my study suggest a possible way of conceptualising both writing and the grading system, and how these might interact with each other. Importantly, features of the writing may determine aspects of its assessment or constrain the ways in which the assessor is able to approach it. Writing at different levels of quality may need to be evaluated in different ways using different criteria or ways of thinking about it. Assessors' possible constructions of the grading system may also exert constraints on how they use it to organise their thinking about a piece of writing. An appreciation of this may make the task seem less mysterious, especially to novice assessors. In addition to this, I have suggested that there may be a distinction to be made between the features that assessors identify in writing and the criteria (both implicit and explicit) they value, and that there may be an interpretive layer between the two. An awareness of this may help to explain difficulties that might be encountered when trying to establish whether writing satisfies certain criteria, and encourage assessors to make their interpretations explicit when sharing their practice with colleagues.

Secondly, I have described six categories of processes that assessors might use in the

course of making an assessment of a piece of discursive writing. An important aspect of my description was that it does not appear that all of these processes are used by all assessors all of the time, or even by one assessor across a whole batch of scripts. Rather the processes consist of a repertoire of analysing and synthesising cognitive behaviours that assessors might recognise themselves using, or that they may consciously wish to draw on, during the course of their assessments. Furthermore, the processes used may be dependent on, or triggered by, the nature of the writing. Awareness of these possible processes may enable assessors to think about how they perform their assessments, and may sensitise them to ways of thinking about their practice that they might not previously have encountered.

The third professional implication relates to a broader insight that the findings may offer for an understanding of the practice of assessment. The variability introduced into assessment by a possible need to evaluate different qualities of writing in different ways; by a need for interpretation of features of writing in relation to criteria (which may themselves be implicit or vary in unsystematic ways); and by different processes being used at different times, or triggered in different ways by the writing being evaluated, may help to explain why there may be inconsistencies across assessments. An awareness of this variability, and its various sources, may help to alter the way in which reliability in assessment is seen, or may need to be seen if the difficulties introduced by inconsistency are to be addressed. It may also suggest that it may be important to re-evaluate the role of reliability in assessment, or to seek alternative ways of thinking about it.

Limitations

A limitation of the study was the inaccessibility of the phenomenon I wished to examine. I sought to gain access to the internal subjective reality of my participants, and I was dependent on their descriptions and my interpretations of their understandings to do so. Participants also sometimes encountered difficulties in verbalising their tacit understandings. However, the difficulty of gaining access to what I needed to

understand became even more apparent when I became conscious that what was implied by the data was something more than frameworks of understanding related to writing and assessment, and a set of procedures the assessors employed to carry out the process of assessment. While these would involve their constructions, it became evident that assessment involved not only frameworks of meaning but was itself an act of construction and interpretation. Thus I was attempting to obtain an understanding of participants' constructions about constructing their views of written work. In addition to a limitation involving my initial simplistic assumptions about the nature of assessment, resulting in the pursuit of data that became less relevant as the study progressed, there was a limitation imposed by the necessary dependence of my findings on several layers of constructions or interpretations. The only means of cross-checking the reality of the criteria and processes distilled from the data was by eliciting similar data from many participants. This was limited by the number of participants it was feasible to interview within the practical and time constraints of the study.

I also asked my participants to talk about their assessment in general, relying on their interpretive synthesis of their practice rather than their description of specific instances. The latter approach would have necessitated a much larger data set in order to be able to cover possible variations in their practice across different instances of assessment. In spite of this attempt to obtain a more varied data set, particularly with respect to processes, not all participants described all of these. In part this may have been related to my focus in the interviews on what appeared salient to each participant. This suggested that a more comprehensive data set and greater saturation might have identified more clearly those processes that are common to all assessors and those that are employed by only a few. It may also have produced fewer instances where interpretations were less strongly supported, and addressed the limited extent to which it was possible to explore the quasi-mathematical integrating processes in complicated assessment and the possible reciprocal relationship between the frameworks of meaning and subjective processes employed by assessors. A more systematic investigation of these might make an additional contribution to an understanding of the assessment process.

A further limitation arising out of the layers of interpretation involved in making sense of the data was the dependence of the research process on my own sense-making. There was also an additional risk attaching to my insider position with respect to the focus of the research and the possibility of this colouring my interpretations. As was noted in the section on methods, the fact that not all of my participants spoke about the same things meant that it was necessary for me to use my overview of the entire data set to go beyond my participants' understandings to build an understanding of the whole. This permitted me to bridge gaps in their descriptions and make sense of the parts. Difficulties encountered by my participants in describing aspects of their practice (suggested by their descriptions not always being very clear, and the misuse of terms) also placed an emphasis on my own interpretations. While my participants were able to comment in follow-up interviews on my interpretations insofar as they dealt with those aspects of their practice that they had talked about, they were more reluctant to confirm or disconfirm my summary descriptions of the apparent emerging structures within the data. The immersion in the entire data set necessary to be able to comment meaningfully on it was beyond the time available in the follow-up interviews. To reduce the possible influence of my position on the data collection, and the possibility of related response bias, I felt that it was important to focus in the interviews on what was salient to my participants. Thus I did not want to force participants to talk about things they had not thought about, which again necessitated a greater reliance on my own interpretations outside the data collection. Although I sought to limit the extent of my subjectivity during the research process, a larger data set would have helped to reduce the extent of the interpretation needed to make sense of the data.

Summary

The account of assessment that I have offered is speculative. An important implication for future research will be to find ways of testing the credibility of this explanation. The description I have offered suggests a conscious systematic processing of information by assessors as they move towards their final assessment of a piece of writing, and I have suggested that a possible thread running through this processing is one of simplification.

I have argued that the participants in my study perform their assessment by means of a deliberative process, although they may not always be able to clearly verbalise all of the constituents of that process. There was no evidence of them employing heuristics or attribute substitution. If this is an accurate account of what assessors do when they assess, the development of a theoretical explanation of the assessment process, grounded in these principles, would permit a greater transparency and a better understanding of the process. Importantly it would imply that the process of assessment does not have to remain obscure, tacit and inaccessible to practitioners and researchers. If it is conscious, it is capable of being influenced consciously, and therefore it can be made open to inspection and improvement. It is likely though that such improvement will be the result of a re-examination of some of the assumptions underpinning current approaches to increasing the reliability of assessment, and the achievement of a theoretical understanding that is considerably more complex than some of the models that have been advanced thus far.

References

Baird J, Greatorex J and Bell J F (2004) 'What makes marking reliable? Experiments with UK examinations' *Assessment in Education* 11 (3) pp 331-348.

Baume D and Yorke M (2002) 'The reliability of assessment by portfolio on a course to develop and accredit teachers in higher education' *Studies in Higher Education* 27 (1) pp 7-25.

Beck S W (2006) 'Subjectivity and intersubjectivity in the teaching and learning of writing' *Research in the Teaching of English* 40 (4) pp 413-460.

Beck S W and Jeffery J V (2007) 'Genres of high-stakes writing assessments and the construct of writing competence' *Assessing Writing* 12 (1) pp 60-79.

Bloxham S (2009) 'Marking and moderation in the UK: False assumptions and wasted resources' *Assessment and Evaluation in Higher Education* 34 (2) pp 209-220.

Bloxham S and Boyd P (2007) *Developing Effective Assessment in Higher Education: A Practical Guide* Maidenhead: Open University Press.

Bloxham S, Boyd P and Orr S (2011) 'Mark my words: The role of assessment criteria in UK higher education grading practice' *Studies in Higher Education* 36 (6) pp 655-670.

Boud D (2007) 'Reframing assessment as if learning were important' in D Boud and N Falchikov (eds) *Rethinking Assessment in Higher Education: Learning for the Longer Term* London: Routledge.

Boud D and Falchikov N (2007) 'Developing assessment for informing judgement' in D Boud and N Falchikov (eds) *Rethinking Assessment in Higher Education: Learning for the Longer Term* London: Routledge.

Bourdieu P (1988) *Homo Academicus* Cambridge: Polity Press.

Broad B (2000) 'Pulling your hair out: Crises of standardization in communal writing assessment' *Research in the Teaching of English* 35 (2) pp 213-260.

Broadfoot P M (1996) *Education, Assessment and Society: A Sociological Analysis* Buckingham: Open University Press.

Calvert B (2005) ‘Demystifying marking: Reflections on developing and using grade descriptors’ *Learning and Teaching in Higher Education* 1 (1) pp 93-97.

Campbell A and Norton L (Eds) (2007) *Learning, Teaching and Assessing in Higher Education: Developing Reflective Practice* Exeter: Learning Matters Ltd.

Charmaz K (2006) *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis* London: Sage Publications.

Cobb P (1994) 'Where is the mind? Constructivist and sociocultural perspectives on mathematical development ' *Educational Researcher* 23 (7) pp 13-20

Condon W (2009) 'Looking beyond judging and ranking: Writing assessment as a generative practice' *Assessing Writing* 14 (3) pp 141-156.

Contreras-McGavin M and Kezar A J (2007) 'Using qualitative methods to assess student learning in higher education' *New Directions for Institutional Research* (136) pp 69-79.

Cooksey R W, Freebody P and Wyatt-Smith C (2007) 'Assessment as Judgment-in-Context: Analysing how teachers evaluate students' writing' *Educational Research and Evaluation* 13 (5) pp 401-434.

Cresswell M J and Houston J G (1991) 'Assessment of the National Curriculum – Some fundamental considerations' *Educational Review* 43 (1) pp 63-78.

Crisp V and Johnson M (2007) 'The use of annotations in examination marking: opening a window into markers' minds' *British Educational Research Journal* 33 (6) pp 943-961.

Crisp V (2008a) 'Exploring the nature of examiner thinking during the process of examination marking' *Cambridge Journal of Education* 38 (2) pp 247-264.

Crisp V (2008b) 'The validity of using verbal protocol analysis to investigate the processes involved in examination marking' *Research in Education* (79) pp 1-12.

Crisp V (2010a) 'Judging the grade: Exploring the judgement processes involved in examination grading decisions' *Evaluation and Research in Education* 23 (1) pp 19-35.

Crisp V (2010b) 'Towards a model of the judgement processes involved in examination marking' *Oxford Review of Education* 36 (1) pp 1-21.

Crisp V (2011) 'The judgement processes involved in the assessment of project work by teachers' *Paper presented at the 12th Conference of the Association for Educational Assessment in Europe, Belfast* [Online] Available http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers [17 Apr 2012].

Crook C, Gross H and Dymott R (2006) 'Assessment relationships in higher education: The tension of process and practice' *British Educational Research Journal* 32 (1) pp 95–114.

Crotty M (1998) *The Foundations of Social Research: Meaning and Perspective in the Research Process* London: Sage Publications.

Cumming A, Kantor R and Powers D E (2002) 'Decision making while rating ESL/EFL writing tasks: A descriptive framework' *The Modern Language Journal* 86 (1) pp 67-96.

Davies M B (2007) *Doing a Successful Research Project: Using Qualitative or Quantitative Methods* Basingstoke: Palgrave MacMillan.

Delaney C M (2005) 'The evaluation of university students' written work' *Paper presented at the British Educational Research Association Annual Conference, Pontypridd* [Online] Available http://www.leeds.ac.uk/educol/documents/149754.htm [10 Apr 2012].

Denscombe M (2007) *The Good Research Guide: For small-scale social research projects (3rd ed)* Maidenhead: Open University Press.

DeRemer M L (1998) 'Writing assessment: Raters' elaboration of the rating task' *Assessing Writing* 5 (1) pp 7-29.

Dreyfus H L and Dreyfus S E (2005) 'Peripheral vision: Expertise in real world contexts' *Organization Studies* 26 (5) pp 779-792.

Ecclestone K (2001) ''I know a 2:1 when I see it': Understanding criteria for degree classifications in franchised university programmes' *Journal of Further and Higher Education* 25 (3) pp 301-313.

Ecclestone K (2002) *Learning Autonomy in Post-16 Education: The Politics and Practice of Formative Assessment* London: Routledge Falmer.

Ecclestone K (2003) *Understanding Assessment and Qualifications in Post-compulsory Education and Training: Principles, Policy and Practice (2nd ed)* Leicester: National Institute of Adult Continuing Education.

Elander J, Harrington K, Norton L, Robinson H and Reddy P (2006) 'Complex skills and academic writing: A review of evidence about the types of learning required to meet core assessment criteria' *Assessment and Evaluation in Higher Education* 31 (1) pp 71-90.

Elbow P (1997) 'Grading student writing: Making it simpler, fairer, clearer' *New Directions for Teaching and Learning (*69) pp 127-140.

Elliott R, Fischer C T and Rennie D L (1999) 'Evolving guidelines for publication of qualitative research studies in psychology and related fields' *British Journal of Clinical Psychology* 38 pp 215-229.

Elwood J and Klenowski V (2002) 'Creating communities of shared practice: The

challenges of assessment use in learning and teaching' *Assessment and Evaluation in Higher Education* 27 (3) pp 243-256.

Ericsson K A and Simon H A (1993) *Protocol Analysis: Verbal Reports As Data (rev ed)* Cambridge MA: MIT Press.

Erling E J and Richardson J T E (2010) 'Measuring the academic skills of university students: Evaluation of a diagnostic procedure' *Assessing Writing* 15 (3) pp 177-193.

Evans J S B T (2003) 'In two minds: dual-process accounts of reasoning' *TRENDS in Cognitive Sciences* 7 (10) pp 454-459.

Evans J S B T (2006) 'Dual system theories of cognition: Some issues' *Proceedings of CogSci 2006, The 28th Annual Conference of the Cognitive Science Society, Vancouver* pp 202-207 [Online] Available http://csjarchive.cogsci.rpi.edu/proceedings/2006/docs/p202.pdf [20 Mar 2011].

Evans J S B T (2007) 'On the resolution of conflict in dual process theories of reasoning' *Thinking and Reasoning* 13 (4) pp 321-339.

Fransella F and Bannister D (1977) *A Manual for Repertory Grid Technique* London: Academic Press.

Funnell E (2000) 'Models of semantic memory' in W Best, K Bryan and J Maxim (eds) *Semantic Processing: Theory and Practice* London: Whurr Publishers.

Gadamer H-G (1993) *Truth and Method (2nd rev ed)* London: Sheed and Ward.

Giddens A (1993) *New Rules of Sociological Method: A Positive Critique of Interpretative Sociologies (2nd ed)* Cambridge: Polity Press.

Gilovich T and Griffin D (2002) 'Introduction – Heuristics and biases: Then and now' in T Gilovich, D Griffin and D Kahneman (eds) *Heuristics and Biases: The Psychology of Intuitive Judgement* Cambridge: Cambridge University press.

Gilovich T, Griffin D and Kahneman D (eds) (2002) *Heuristics and Biases: The Psychology of Intuitive Judgement* Cambridge: Cambridge University press.

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

Glaser B G (1992) *Basics of Grounded Theory Analysis* Mill Valley CA: Sociology Press.

Grainger P, Purnell K and Zipf R (2008) 'Judging quality through substantive conversations between markers' *Assessment and Evaluation in Higher Education* 33 (2) pp 133-142.

Greatorex J (1999) ‘Generic descriptors: A health check’ *Quality in Higher Education* 5 (2) pp155-166.

Greatorex J (2001) ‘Making the grade – How question choice and type affect the development of grade descriptors’ *Educational Studies* 27 (4) pp 451-464.

Greatorex J (2002) ‘Making accounting examiners’ tacit knowledge more explicit: Developing grade descriptors for an accounting A-level’ *Research Papers in Education* 17 (2) pp 211-226.

Greatorex J (2007) 'What strategies do IGCSE examiners use to mark candidates' scripts?' *International Schools Journal* 27 (1) pp 48-55.

Greatorex J, Johnson C and Frame K (2001) ‘Making the grade – Developing grade descriptors for accounting using a discriminator model of performance’ *Westminster Studies in Education* 24 (2) pp 167-181.

Greatorex J and 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* [Online] Available http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/194453_IAEA_2006.pdf [02 Sep 2010]

Haggis T (2006) 'Pedagogies for diversity: retaining critical challenge amidst fears of 'dumbing down'' *Studies in Higher Education* 31 (5) pp 521-535.

Haines C (2004) *Assessing Students’ Written Work: Marking Essays and Reports* London: RoutledgeFalmer.

Hamp-Lyons L (1991) ‘Pre-text: Task-related influences on the writer’ in L Hamp-Lyons (ed) *Assessing Second Language Writing in Academic Contexts* Norwood: Ablex Publishing Corporation.

Hoyt J E, Allred E R and Hunt R (2010) 'Implementing writing assessment in a degree completion program: Key issues and lessons learned' *The Journal of Continuing Higher Education* 58 pp 19-30.

Hunter K and Docherty P (2011) 'Reducing variation in the assessment of student writing' *Assessment and Evaluation in Higher education* 36 (1) pp 109-124.

Huot B (1990) ‘Reliability, validity, and holistic scoring: What we know and what we need to know’ *College Composition and Communication* 41 (2) pp 201-213.

Huot B (1996) ‘Toward a new theory of writing assessment’ *College Composition and Communication* 47 (4) pp 549-566.

Huot B (2002) *(Re) Articulating Writing Assessment for Teaching and Learning* Logan: Utah State University Press.

Huot B and Perry J (2009) 'Toward a new understanding for classroom writing assessment' in R Beard, D Myhill, J Riley and M Nystrand (eds) *The Sage Handbook of Writing Development* London: Sage Publications.

Isaacs T (2010) 'Profiles of education assessment systems worldwide: Educational assessment in England' *Assessment in Education: Principles, Policy and Practice* 17 (3) pp 315-334.

Jawitz J (2009) 'Learning in the academic workplace: The harmonization of the collective and the individual habitus' *Studies in Higher Education* 34 (6) pp 601-614.

Jeffery J V (2009) 'Constructs of writing proficiency in US state and national writing assessments: Exploring variability' *Assessing Writing* 14 (1) pp 3-24.

Johnson M and Greatorex J (2005) 'Judging learners' work on screen. How valid and fair are assessment judgements?' *Paper presented at the British Educational Research Association Annual Conference, Pontypridd* [Online] Available http://www.leeds.ac.uk/educol/documents/150295.htm [02 Sep 2010].

Johnston B (2004) 'Summative assessment of portfolios: An examination of different approaches to agreement over outcomes' *Studies in Higher Education* 29 (3) pp 395- 412.

Kahneman D and Frederick (2002) 'Representativeness revisited: Attribute substitution in intuitive judgement' in T Gilovich, D Griffin and D Kahneman (eds) *Heuristics and Biases: The Psychology of Intuitive Judgement* Cambridge: Cambridge University press.

Kahneman D, Slovic P and Tversky A (eds) (1982) *Judgement Under Uncertainty: Heuristics and Biases* Cambridge: Cambridge University Press.

Klein G A (1997) 'The recognition-primed decision (RPD) model: Looking back, looking forward' in C E Zsambok and G A Klein (eds) *Naturalistic Decision Making* Mahwah NJ: Lawrence Erlbaum Associates.

Knight P T (2002) 'Summative assessment in higher education: Practices in disarray ' *Studies in Higher Education* 27 (3) pp 275-286.

Knight P (2006) 'The local practices of assessment' *Assessment and Evaluation in Higher Education* 31 (4) pp 435-452.

Kreth M, Crawford M A, Taylor M and Brockman E (2010) 'Situated assessment: Limitations and Promise' *Assessing Writing* 15 (1) pp 40-59.

Kvale S (1996) *Interviews: An Introduction to Qualitative research Interviewing* London: Sage Publications.

Laming D (2004) *Human Judgement: The Eye of the Beholder* London: Thomson.

Langdridge D (2007) *Phenomenological Psychology: Theory, Research and Method* Harlow: Pearson Education Ltd.

Laverty S M (2003) 'Hermeneutic phenomenology and phenomenology: A comparison of historical and methodological considerations' *International Journal of Qualitative Methods* 2 (3) pp 1-29.

Lea M R and Street B V (1998) 'Student writing in higher education: An academic literacies approach' *Studies in Higher Education* 23 (2) pp 157-172.

Lincoln Y S and Guba E G (1985) *Naturalistic Inquiry* London: Sage Publications.

Lincoln Y S and Guba E G (2000) 'Paradigmatic controversies, contradictions, and emerging confluences' in N K Denzin and Y S Lincoln (eds) *Handbook of Qualitative Research (2nd ed)* London: Sage Publications.

Lumley T (2002) 'Assessment criteria in a large-scale writing test: What do they really mean to the raters?' *Language Testing* 19 (3) pp 246-276.

Lyotard J-F (1991) *Phenomenology* Translated by Beakley B, Albany NY: State University of New York Press.

Marshall B (2007) 'Assessment in English' First published in T Cremin and H Dombey (eds) *Handbook of Primary English in Initial Teacher Education* [Online] Available http://www.ite.org.uk/ite_readings/index.html [10 Sep 2009].

Milanovic M, Saville N and Shuhong S (1996) 'A study of the decision-making behaviour of composition markers' in M Milanovic and N Saville (eds) *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium, Cambridge* Cambridge: Cambridge University Press.

Moran D (2000) *Introduction to Phenomenology* London: Routledge.

Moustakas C (1994) *Phenomenological Research Methods* London: Sage Publications.

Mullins G and Kiley M (2002) 'It's a PhD, not a Nobel Prize: How experienced examiners assess research theses' *Studies in Higher Education* 27 (4) pp 369-386.

Naidoo R and Jamieson I (2005) 'Empowering participants or corroding learning? Towards a research agenda on the impact of student consumerism in higher education' *Journal of Education Policy* 20 (3) pp 267-281.

O'Donovan B, Price M and Rust C (2004) 'Know what I mean? Enhancing student understanding of assessment standards and criteria' *Teaching in Higher Education* 9 (3) pp 325-335.

Osman M (2004) 'An evaluation of dual-process theories of reasoning' *Psychonomic Bulletin and Review* 11 (6) pp 988-1010.

Polit D F and Beck C T (2006) *Essentials of Nursing Research: Methods, Appraisal, and Utilization (6th ed)* London: Lippincott Williams and Wilkins.

Pollitt A and Murray N L (1996) 'What raters really pay attention to' in M Milanovic and N Saville (eds) *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium, Cambridge* Cambridge: Cambridge University Press.

Price M (2005) 'Assessment standards: The role of communities of practice and the scholarship of assessment' *Assessment and Evaluation in Higher Education* 30 (3) pp 215-230.

Price M, O'Donovan B, Rust C and Carroll J (2008) 'Assessment standards: A manifesto for change' *Brookes eJournal of Learning and Teaching* 2 (3) [Online] Available http://bejlt.brookes.ac.uk/article/assessment_standards_a_manifesto_for_change/ [26 Feb 2012]

Read B, Francis B and Robson J (2004) 'Re-viewing undergraduate writing: Tutors' perceptions of essay qualities according to gender' *Research in Post-Compulsory Education* 9 (2) pp 217-238.

Read B, Francis B and Robson J (2005) 'Gender, 'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays' *Assessment and Evaluation in Higher Education* 30 (3) pp 241-260.

Rezaei A R and Lovorn M (2010) 'Reliability and validity of rubrics for assessment through writing ' *Assessing Writing* 15 (1) pp 18-39.

Ricoeur P (1970) *Freud and Philosophy: An Essay on Interpretation* Translated by Savage D, New Haven: Yale University Press.

Ricoeur P (1981) *Hermeneutics and the Human Sciences: Essays on Language, Action and Interpretation* Translated by Thompson J B, Cambridge: Cambridge University Press.

Rogoff B (1990) *Apprenticeship in Thinking: Cognitive Development in Social Context* Oxford: Oxford University Press.

Rust C (2007) 'Towards a scholarship of assessment' *Assessment and Evaluation in Higher Education* 32 (2) pp 229-237.

Rust C (2011) 'The unscholarly use of numbers in our assessment practices: What will make us change?' *International Journal for the Scholarship of Teaching and Learning* 5 (1) pp 1-6.

Sadler D R (1989) 'Formative assessment and the design of instructional systems' *Instructional Science* 18 (2) pp 119-144.

Sadler D R (2010) 'Fidelity as a precondition for integrity in grading academic achievement' *Assessment and Evaluation in Higher Education* 35 (6) pp 727-743.

Sanderson P J (2001) 'Language and Differentiation in Examining at A level' *Unpublished PhD thesis* University of Leeds [Online] Available http://lib.leeds.ac.uk/record=b2244065 [07 Nov 2010]

Schmidt L K (2006) *Understanding Hermeneutics* Stocksfield: Acumen Publishing Ltd.

Schwandt T A (2000) 'Three epistemological stances for qualitative inquiry: Interpretivism, hermeneutics, and social constructionism' in N K Denzin and Y S Lincoln (eds) *Handbook of Qualitative Research (2nd ed)* London: Sage Publications.

Scott M (2005) 'Student writing, assessment, and the motivated sign: Finding a theory for the times' *Assessment and Evaluation in Higher Education* 30 (3) pp 297-305.

Scott T (2008) "Happy to comply': Writing assessment, fast capitalism, and the cultural logic of control' *The Review of Education, Pedagogy, and Cultural Studies* 30 pp 140-161.

Shay S B (2004) 'The assessment of complex performance: A socially situated interpretive act' *Harvard Educational Review* 74 (3) pp 307-329.

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

Shay S (2008) 'Assessment at the boundaries: Service learning as case study' *British Educational Research Journal* 34 (4) pp 525-540.

Shute V J, Torreano L A and Willis R E (2000) 'DNA: Providing the Blueprint for Instruction' in J M Schraagen, S F Chipman and V L Shalin (eds) *Cognitive Task Analysis* Mahwah: Lawrence Erlbaum Associates.

Sloman S A (2002) 'Two systems of reasoning' in T Gilovich, D Griffin and D Kahneman (eds) *Heuristics and Biases: The Psychology of Intuitive Judgement* Cambridge: Cambridge University press.

Slomp D H (2012) 'Challenges in assessing the development of writing ability: Theories, constructs and methods' *Assessing Writing* 17 (2) pp 81-91.

Stanovich K E and West R F (2002) 'Individual differences in reasoning: Implications for the rationality debate?' in T Gilovich, D Griffin and D Kahneman (eds) *Heuristics and Biases: The Psychology of Intuitive Judgement* Cambridge: Cambridge University press.

Stanovich K E, Toplak M E and West (2008) 'The development of rational thought: A taxonomy of heuristics and biases' in R V Kail (ed) *Advances in Child Development and Behaviour Vol 36* London: Elsevier.

Stiles W B (1993) 'Quality control in qualitative research' *Clinical Psychology Review* 13 (6) pp 593-618.

Stobart G (2008) *Testing Times: The Uses and Abuses of Assessment* Abingdon: Routledge.

Strauss A and Corbin J (1998) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* London: Sage Publications.

Suto W M I and Greatorex J (2008a) 'A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations' *Assessment in Education: Principles, Policy and Practice* 15 (1) pp 73-89.

Suto W M I and Greatorex J (2008b) 'What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process' *British Educational Research Journal* 34 (2) pp 213-233.

Tobin G A and Begley C M (2004) 'Methodological rigour within a qualitative framework' *Journal of Advanced Nursing* 48 (4) pp 388-396.

Tversky A and Kahneman D (1974) 'Judgement Under Uncertainty: Heuristics and Biases' *Science* 185 pp 1124-1131 reprinted in T Connolly, H R Arkes and K R Hammond (eds) (2000) *Judgement and Decision Making: An Interdisciplinary Reader (2nd ed)* Cambridge: Cambridge University Press.

van Manen M (1990) *Researching Lived Experience: Human Science for an Action Sensitive Pedagogy* Albany NY: State University of New York Press.

van Merrienboer J J G, Kirschner P A and Kester learning (2003) 'Taking the load off a learner's mind: Instructional design for complex learning ' *Educational Psychologist* 38 (1) pp 5-13.

Vaughan C (1991) 'Holistic assessment: What goes on in the raters' minds' in L Hamp-Lyons (ed) *Assessing Second Language Writing in Academic Contexts* Norwood: Ablex Publishing Corporation.

Von Glasersfeld E (1989) 'Cognition, Construction of Knowledge, and Teaching ' *Synthese* 80 (1) pp 121-140

Walsh P (2006) 'Narrowed horizons and the impoverishment of educational discourse: Teaching, learning and performing under the new educational bureaucracies' *Journal of Education Policy* 21 (1) pp 95-117.

Webster F, Pepper D and Jenkins A (2000) 'Assessing the undergraduate dissertation' *Assessment and Evaluation in Higher Education* 25 (1) pp 71-80.

Weigle S C (2002) *Assessing Writing* Cambridge: Cambridge University Press.

Wenger E (1998) *Communities of Practice: Learning, Meaning, and Identity* Cambridge: Cambridge University Press.

Willig C (2001) *Introducing Qualitative Research in Psychology: Adventures in Theory and Method* Maidenhead: Open University Press.

Wilson V (1997) 'Focus Groups: A useful qualitative method for educational research?' *British Educational Research Journal* 23 (2) pp 209-224.

Wolcott W (1998) *An Overview of Writing Assessment* Urbana IL: National Council of Teachers of English.

Wolf A (1995) *Competence-Based Assessment* Buckingham: Open University Press.

Woolf H (2004) 'Assessment criteria: Reflections on current practices' *Assessment and Evaluation in Higher Education* 29 (4) pp 479-493.

Wyatt-Smith C and Castleton G (2005) 'Examining how teachers judge student writing: An Australian case study' *Journal of Curriculum Studies* 37 (2) pp 131-154.

Wyatt-Smith C, Klenowski V and Gunn S (2010) 'The centrality of teachers' judgement practice in assessment: A study of standards in moderation' *Assessment in Education: Principles, Policy and Practice* 17 (1) pp 59-75.

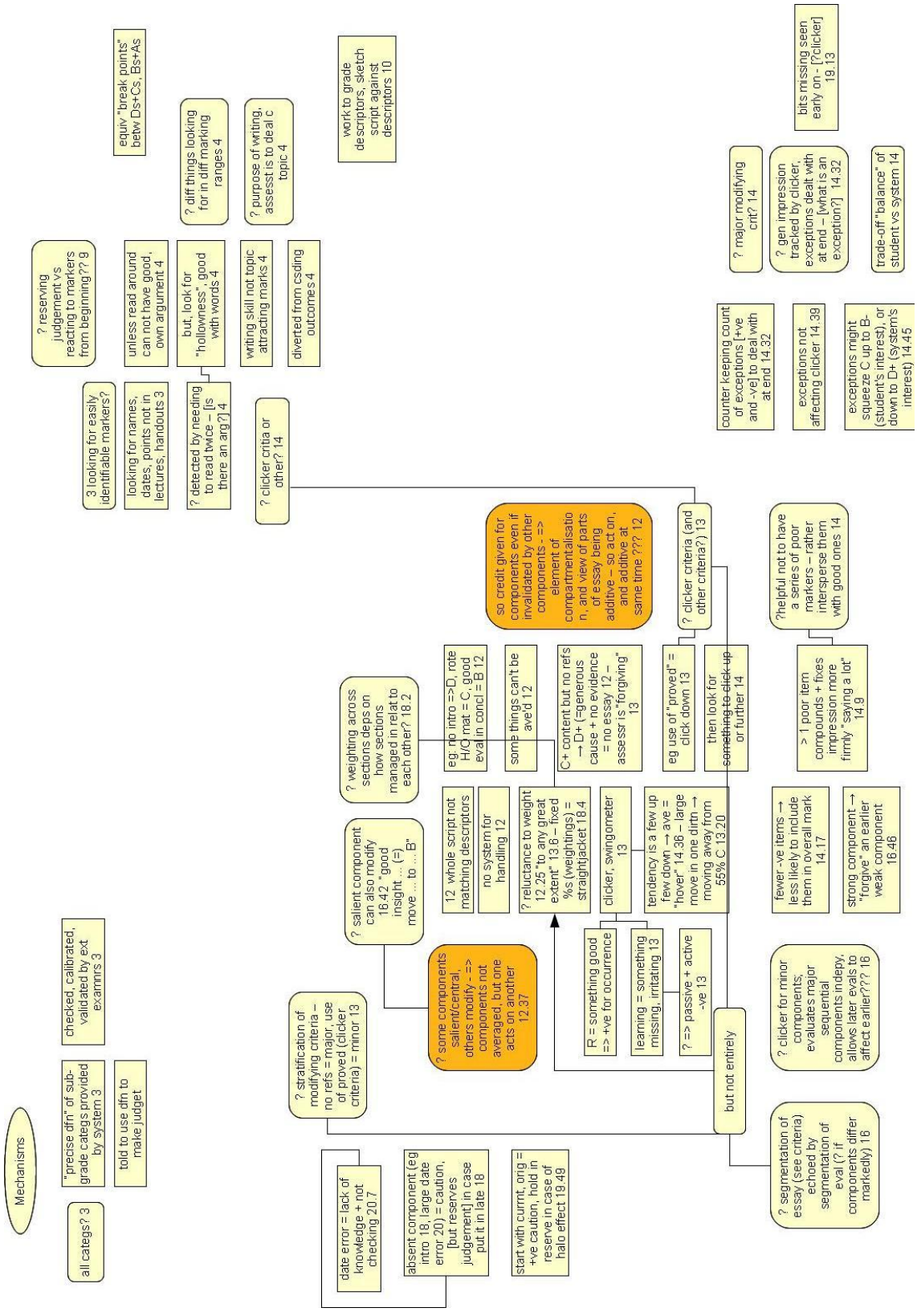
Yorke M (2008) *Grading student achievement in higher education: Signals and shortcomings* London: Routledge.

Young R A and Collin A (1988) 'Career development and hermeneutical inquiry Part I: The framework of a hermeneutical approach' *Canadian Journal of Counselling* 22 (3) pp 153-161.

Appendix 1: Interview questions

Questions/topics of interest to the researcher	**Probe questions/topics for participants** *(considering both the perspective of 'an assessor' and the participant as an assessor her/himself)*
Purpose of assessment	*Thinking generally about assessment ...*
	What comes to mind?
	Purpose of assessment?
	Role/position/status of assessor and student in the process? (Aims of assessment, use made of assessment)
	Are you a typical assessor?
Purpose of writing	*I'm mainly interested in the assessment of writing ...*
	Use/purpose/meaning/function of writing? (HE context?)
	How is (your) assessment related to that purpose?
	What are you/(is being) focussed on in assessment?
	What kind of (academic?) competence might be shown through writing?
Procedure/mechanics of assessment	*Thinking about the procedure, reading and evaluating a piece of writing ...*
	What is your conceptualisation of the process? General principles, or you as assessor, or from student experiences – whatever is salient
What does assessor <u>do</u>	If you have a particular style/technique/method for reading/handling essays, can you describe it? (Steps/stages/strategy/focus) Reflect on its utility?
Perception of quality	How do you construct your perception of the quality of the work? (What 'grabs' you, or puts you off?) (Meaning of argument, logic, flow, structure, coherence?) (Role of your initial impression?) What is the use of rubric/guidelines in this? (How do you deal with the rubric profile?)
Grade bands,	Where does the grade come from? (Sources of grades?) Meanings of grade bands?
Borderline *Differentiating* *Scaling*	What about grade boundaries, any salience of the borderline? Differentiating between marks (grades, semi-grades, %)? Using the top of the range, scaling of judgement?
Good versus poor writing *Consistency of writing*	How do you approach a good versus a poor piece of writing? How do you approach a consistent versus an irregular profile?
Overall	*Thinking about our discussion today ...*
	Any other thoughts, things we perhaps haven't talked about that might be important?

Appendix 2: Example of data analysis - Drawing display



Appendix 3: Example of data analysis - Spreadsheet display

		? overlap w qualified scale]		ok to be one grade range 17.33	A+	outstanding	wow	v rare, "reserved" 13	=creativity	not = student, publishable quality ito depth, amount, maturity of style 11.19	diff to other students – hence form of creative 12; hence novel, unusual [one off?]	doesn't have to be creative	(cf res paper v good job, no glaring omissions, skilfully interpreted	[? ave published paper]	strucd as a textbook would – flow is as predicted [? easy to read/predict]	OR = creativity – use journs not mentioned, diff way of looking diff order [?alt perspective]
marking limit			need 80% grade		A		depth, clarity									
transition more imp 22			7% jump – need A-- = 70, 71	need more banding 17.33	A-		depth, clarity		difference, special 9 not necessary earth shattering) 11.19	?stands out from others	Can't get A- (and above) on content alone – reqs style, flow					
rarer, special, scrutinised more		can differentiate, discrim 21.48			B+	clarity, orig of arg/thought procs, diff perspective										
					B		extra material from wider reading	content building, increasing depth or sophistication, ? stands out from others								
				v imp for most students	B-		increasing content 3, 11			basic mat + smooth style, good intro/concl						
mean grade 22.7 [? mode]	[? majority of essays]			transition less imp – but should not be so 22	C+	textbook, "bluff"	content building 12	comp, but not enough for B- - ? content hasn't grown yet21.11	[? comp = no mess ups] 21.11	comp basic mat						
					C	="comp" [22.13]		basic ans, a few mess ups								
		diffic to discrim 21.11			C-			basic plus a few mistakes								
			?ok to be one band 21.11 =D-	diffic to discrim 20	D+	diffic to discrim, differentiate	diffic to discrim21.5, 21.48									
					D	20.20										
	[? fewer read]				D-		done "just enough"21.5									
				not too diffic to discrim	F	written nothing, or too short, or longer but it's rubbish										

Appendix 4: Example of data analysis - Narrative summary

Source of the grade:

p23
In describing how the grade "emerges", which might be interpreted to suggest that the grade is viewed as being a property of the work rather than being constructed by the assessor, P4 described his "first thought" as being that the work is "a 2:1, a 2:2 or a first". He suggested that he was "probably getting a good idea of where I think it's going to go" by halfway through the essay, and that this "feel" would be developing with his reading (i.e. he did not suspend his judgement until some point in the reading of the essay). Further he thought that his impression could start within the first half to one page and that this would be a function of the "style" of the writing and the "quality" of the language used - "you can just tell there's some quality there from the start, the language used ... well written English basically and (it) just has a maturity, a sort of elegance about it I suppose that you don't often find in a lot of work". He suggested that for most of the cases the writing does not jar but that " occasionally something will jump out at you". He agreed that writing might display qualities of "jarring", "not jarring", or "jumping out at you". He felt that the introductory paragraph could "set the scene", and although it is possible to "recover" the impression of the work builds as the writing develops such that the impression becomes increasingly more established the further the assessor reads. He felt that by half way the assessor had begun to "build up a picture". Although it is still "definitely not game over" at that stage, a piece of writing can become more difficult to retrieve or "screw up".

p27
At the end of the essay the assessor has got a feel for the overall grade (2:1, first). In arriving at a final grade the participant said that the criteria "tick boxes don't allow me to say the grade" (29.13), suggesting that his impression of the work is constructed more holistically. At this stage he will go through the marking criteria he has set, sometimes accompanied by a "re-flick" through the essay. This may prompt him to recognise a mismatch between his holistic feeling about the essay and his criteria. This mismatch may be resolved by an adjustment to the holistic impression, or by considering the weightings he wishes or feels he ought to assign to his criteria. My impression is that the criteria assist in the interrogation of the rationale for the holistic grade, or that this process might provoke a refinement of the criteria. The criteria act as a means for testing or verifying the holistic grade, or modifications to that grade, rather than there being two means of arriving at the grade that need to be reconciled (29.13). The criteria might also help the assessor to work out why he has arrived at a particular mark or grade and are a part of the rationalising process. The criteria (in the case of this participant) are not applied mechanistically, but serve to provoke thinking in the assessor regarding his impression and whether he might need to revise his impression. Going through the criteria feeds into the assessor's thinking about the mark.

p30
In thinking about variations in consistency across a piece of writing, he suggested that he would tend to average out the good and poorer bits, and that a very good section in a piece of writing would not over-ride poorer bits. He valued consistency, and felt that to get a top mark all parts of an essay would need to be equally excellent. The participant felt that he would view more favourably something that started off poorly but finished well, whereas he would view a good start that finished poorly as demonstrating carelessness. He saw his approach to the process of marking as involving both rewarding and punishing, but that he did "like to reward" when a student has done something different or stood out. He agreed that he tended to look more for things to reward that stood out from a baseline than to punish for falling below that baseline, and that this reward would be directed towards the writing that had "done the extra things".