EDUCATION-LINE

Grading Student Work: Using think aloud to investigate the assessment practices of university lecturers

Pete Boyd* University of Cumbria
Mary Ashworth University of Cumbria
Sue Bloxham University of Cumbria
Susan Orr York St john University

*contact for correspondence: pete.boyd@cumbria.ac.uk

Paper presented at the British Educational Research Association Annual Conference, University of Manchester, 2-5 September 2009

Marking is important. The grades we give students and the decisions we make about whether they pass or fail coursework and exams are at the heart of our academic standards. 'Markers are the gatekeepers for university quality' (Smith & Coombe, 2006:45). Assessment in higher education involves highly decentralised and subject-specific decision-making given credence in the UK context by processes of quality assurance ranging from national agencies and review systems to local teaching team practices in marking and moderation. This quality assurance is currently firmly located within a paradigm of accountability, explicit learning outcomes, constructive alignment (Biggs & Tang, 2007), transparency and criteria-based marking (Quality Assurance Authority, 2006). Whilst much of the assessment system is placed in the public domain, subject to scrutiny and debate, there remains considerable mystery surrounding the marking activity by lecturers that underpins the whole apparatus (Yorke, Bridges & Woolf, 2000).

In the assessment of open-ended coursework assignments in higher education Sadler (2009a) identifies anomalies in the apparently systematic appraisal process when using explicit criteria and argues that holistic appraisals using expert judgement offer considerable potential and deserve more rigorous empirical study. This pilot study uses think aloud protocols, asking assessors to verbalise their thinking, to investigate the marking practices of university lecturers as they grade and provide feedback on written coursework assignments. It asks the question, how do lecturers make judgements about student work and what is the role of assessment criteria within that process? The use of think aloud on 'live' marking of student work presents some ethical risks but is intended to shine some light into the mystery of university marking. This paper presents the emerging findings of the pilot project and begins to evaluate the think aloud method as an approach to investigating the marking practice of university lecturers.

The practice of marking

Higher education marking in the UK takes place within a quality assurance framework of accountability with a focus on transparency and criteria-based marking. As Grainger, Purnell and

Zipf point out (2008), this pressure for accountability requires assessment decisions to be justified and this is reflected in the Quality Assurance Agency (QAA) Code of Practice which specifies the requirement that 'Institutions have transparent and fair mechanisms for marking and moderation' (QAA, 2006:16). These 'validating practices' (Shay 2004) for assessment are intended to make the processes and judgements of assessment more transparent to staff and students and to reduce the arbitrariness of staff decisions (Sadler 2009a).

One result is that 'production, publication and discussion of clear assessment criteria are a sine qua non of an effective assessment strategy' (Woolf 2004:479) and, as Sadler (2009a) states, using criteria is considered best practice to the point that they are mandatory in some Universities. They have come into widespread use, according to Sadler because of the benefits they offer in terms of ethical practice, providing guidance, greater objectivity in marking, and communicating feedback more easily. There is also some evidence that they can make a difference to student learning (Bloxham & Boyd, 2007). It has to be said that there is some confusion between criteria and standards (Grainger et al, 2008) and they are probably used interchangeably by different people and in different contexts.

Sadler (2005) distinguishes *criteria*, that are designed to judge how well the student has demonstrated progress towards the desired learning outcomes, from *standards,* which specify qualitative criteria or attributes required. In this analysis *standards* will often be seen as a physical artefact in the typical UK departmental or university wide grade descriptors which specify what students must do in order to achieve a particular grade. This distinguishes *criteria* as likely to be specific to a given assignment whereas *standards* might apply across all work at the relevant level.

Despite the emphasis placed on the use of assessment criteria there is a lack of theoretical or research support that 'one might have assumed' (Sadler, 2008). A survey of the literature indicates three key sources of criticism of this paradigm of accountability in relation to criterion-referenced assessment and marking,

Socio-cultural critique

Orr (2007) argues that 'validating practices' for assessment in higher education are based on a techno-rationalist approach to thinking about assessment and a positivist model of assessment standards. Delandshere notes that assessment practice is based on assumptions that 'knowledge is monolithic, static and universal' (2001: 127), a view echoed by Shay (2004). Shay draws on a range of theoretical accounts in discussing how the implicit rationality in Western values hides a rationality which is rather 'a context-dependent, experience-based, and situational judgement' (p323). However, a techno-rational view of knowledge underpins the accountability paradigm. It equates publication with explicitness and, as Gonzalez Arnal & Burwood (2003) argue, this does not stand up to scrutiny because it:

> ...is based on a model of knowledge that ought to be resisted and that is, at its core, false. Assessment consists in the exercise of an applied skill, and there are core aspects of this knowledge practice that *cannot be captured by a mere propositional description* of them, thus making them unavailable for publication' (2003:382, emphasis from the original)

A socio-cultural view also recognizes the private and tacit nature of assessment knowledge. And whilst, they are not attempting to make a sociological or political point, other researchers challenge

the notion that it is possible to make explicit the tacit knowledge involved in assessment decisions (Sadler, 2009a; Orr, 2007; Shay, 2005; O'Donovan, Price & Rust, 2008).

The 'hidden' and inexpressible nature of this tacit knowledge is compounded by the complex nature of work being assessed at higher education level which allows for a wide range of satisfactory student responses. For example, students may respond to an essay question or design brief in very different, but equally effective, ways. This requires tutors to use their judgement, based on their tacit knowledge, in order to allocate grades. Eisner (1985) refers to this process as the use of 'connoisseurship'; the well-informed subjective judgment which accrues through immersion in a subject discipline. It is an 'interpretivist' view of assessment which recognizes the power of the local context (Elton & Johnston 2002; Knight & Yorke, 2003). Indeed Shay (2004: 309) describes higher education assessment as a 'socially situated interpretive act' and Stowell (2004), in discussing equality in higher education, reinforces this view in arguing that 'in reality what constitutes merit or academic achievement is a social decision and a product of social relations' (p 498).

However, Shay (2005) argues that although such judgement is subjective at one level, it gains objectivity from being informed by the tacit standards, norms and rules of the particular academic field and allows for an element of professional and local interpretation even though there is evidence that inconsistency in marking exists (Bloxham, 2009). From this perspective, written assessment criteria have limited power to secure national standards as their interpretation will be determined locally (Knight, 2006) by tutors drawing on their experience and therefore their differing tacit knowledge of disciplinary standards (Knight, Yorke 2003, Price, Rust 1999, Ecclestone, 2001). Varying professional knowledge, experience and values lead staff to attach importance to different qualities in student work (Smith & Coombe, 2006; Read, Francis & Robson, 2005).

Our 'common sense' response to concerns about these potential variations in assessors practices and standards has been to publish explicit assessment criteria and standards for coursework but this is undermined by the difficulty of communicating their meaning (Price & Rust, 1999; Ecclestone, 2001), and by tutors' customary approaches to marking. According to Wolf (1995), markers may acquire fixed habits in their marking, which they are unaware of, but which can influence their grading. Staff develop predispositions and biases (Grainger et al, 2008) and assessors may not understand or agree with outcomes they are supposed to judge (Baume, Yorke & Coffey, 2004). Evidence also suggests that staff ignore criteria, choose not to adopt them (Smith & Coombe, 2006; Price & Rust, 1999; Ecclestone, 2001) or use implicit standards which may contradict published standards (Price, 2005; Read, Francis & Robson, 2005; Baume, Yorke & Coffey, 2004).

From a socio-cultural perspective marking within a teaching team can be seen as a collective dynamic object-oriented activity system (Engestrom, 1987, 2001) in which formal and informal 'rules' govern tutor behaviour. The tutors' professional knowledge as applied to the marking activity is mediated, situated, social, provisional and contested (Blackler, 1995) and such assessment must be understood as a socio-political practice. In examining the historical place of assessment in society and examining the assumptions underpinning contemporary assessment practice it can be argued that generally the systems in place in western economies play a role in reinforcing social structures (Delandshere, 2001).

While Delandshere sees the significant historical and cultural context as being at a wider scale, for example in the higher education context, previous work on professional learning of academics has identified the subject department as the critical unit of analysis (Knight & Trowler, 2001). At this

middle level of analysis as activity within a subject department or teaching team, between the individual and the wider context, analysis of assessment practice from a socio-cultural perspective has examined the social nature of grading and moderation (Orr, 2007; 2008). In her study of practice in a UK art department Orr found of examples of grade decisions being affected by power relationships within teams and examples where knowledge of the student appeared to lead to an ipsative approach to marking, meaning that the previous effort and achievement of the student and the motivational effect of a particular grade were considered by lecturers. From an activity systems theoretical perspective (Engestrom, 1987) the wider historical and cultural contextual influences may be revealed through analysis of the 'rules', the formal and informal constraints on practice, that are applied by the tutors in the marking activity.

Theoretical critique

The 'accountability paradigm' as reflected in assessment and marking is based on certain assumptions. Firstly, it assumes that criterion-referenced assessment is possible without reference to student norms. Secondly, it assumes that we can write unambiguous statements of criteria or standards which can be consistently interpreted by students and staff (Tan & Prosser, 2004) across highly unstandardised assessment tasks. Thirdly, there is an assumption that we can allocate different ranks of marks (standards) across a range of criteria in a reliable way. Finally it assumes that we are able to mentally manipulate a complex set of explicit criteria whilst reading student work in order to make a grading decision, this is what Sadler refers to as analytic grading (2009a).

The 'accountability paradigm' as expressed in most guidance to staff works from the basis that we should base our assessment on criterion rather than norm referencing so that a student is judged against a set of standards, not against his or her peers. This distinction has been criticised (Neil, Wadley 1999) and reasons for the blurring of the distinction in practice, despite the clarity of the distinction in theory, is discussed by Yorke (2009). In particular, he makes the point that 'assessors' grading behaviour is tacitly influenced by norm referencing' (p69) and Shay (2004) also found that tutors draw on their knowledge of different students work in order to make their judgements (p320).

A further pressure on marking consistency is the open and diverse nature of student work. Shay (2004) argues that, traditionally, reliability has been based on highly standardized assessment tasks, whereas higher education assessment (in the case of her paper – final year projects) is characterised by low levels of standardisation and responses to the task can be diverse in terms of the skills the students use. These complex tasks create problems for inter-marker reliability, and although there is a lot of common ground in marking, there is also considerable disparity.

Criterion or standards-based assessment presupposes an analytical approach to grading; that is that the marker makes separate qualitative judgements on a range of preset criteria (Sadler 2009a). However, Sadler makes a strong case for the view that an 'analytic approach is theoretically and practically deficient on two grounds. By limiting itself to preset criteria, it cannot take into account all the necessary nuances of expert judgements. Neither can analytic appraisal, when using a simplistic combination rule, represent the complex ways in which criteria are actually used…..the 'truer' representation is the 'fuller' of the two' (p177). In other words, he is challenging the view that lecturers assess criteria by criteria and that, in practice, they decide on the respective contribution of different criteria after they have made an assessment of the work. He draws on the notion of complex judgements to argue that criteria often merge in practice or interact with each other and staff use different sets of criteria for the same type of assignment (e.g. an essay). In addition, he

makes the point (p176) that the critique of subjectivity of the holistic approach to assessment could equally well be applied to the subjectivity inherent in the application of each individual criterion. Overall he urges a greater acknowledgement of the contribution of holistic judgement.

Sadler's views that pre-determined criteria do not reflect the full range of criteria being used to judge students' work is echoed in Yorke's (2009) work and in the third of Tan and Prossers (2004) conceptions of grade descriptors (grade indicators) where assessors consider that they do not 'depict the full range of desired qualities of student work'.

Empirical studies

Shay (2004: 315) draws on Bourdieu (1988) to discuss how systems of classification in higher education (including systems used in assessment) are never codified, but are 'subconscious, acquired through practical mastery'. Interestingly, whilst Sadler's arguments may be based in a technical analysis of criteria and Shay's in a sociological discourse, the conclusions are the same, we do not have transparency in assessment judgements. We mislead students that there is something fixed, accessible and rational that they can use to guide their work.

This doesn't mean that staff don't feel confident getting on with making judgements. Over time they acquire the practical mastery which guides their decisions, and which is a form of socialization that emerges from doing the job, marking over and over again. People learn to mark by marking. Consequently Shay found that assessors were drawing on this 'feel for the game' (Bourdieu & Wacquant, 1992) rather than explicit 'marking memoranda' (assessment criteria or standards) in marking.

Shay's respondents also identified the practical issues associated with trying to use written criteria, struggling with the 'false compartmentalisation' (316), the difficulty of manipulating a range of criteria simultaneously and the problem of trying to articulate how they are assessing. Staff claimed instead to use an holistic approach to making judgements. Not surprisingly, then, she did find marking differences as assessors judgements emerged from different academic specialisms, value systems, involvement with students and experience.

Orr & Blythman (2005) have explored the disjuncture that can exist between written guidance regarding assessment and how it is done in practice and Orrell (2008) found differences between tutors' espoused beliefs about marking and their actual marking practice. In Orrell's study there was little espoused concern about using assessment criteria or other ways of achieving accuracy and consistency in their grading. Furthermore, tutors' actual marking practice did not reveal the use of either technical strategies or 'qualitative measures that would either improve the grading reliability or give explicit meaning to their grades' (p198).

Yorke (2009) also discusses holistic versus analytic grading. He argues that the two different approaches can create very different results. However, he suggests that in practice, people don't do one or the other, but probably markers have a tendency towards one approach or the other – "there is some evidence that assessors juggle their marking until they arrive at a mark or grade with which they feel comfortable" (p69). This is supported in the work of Grainger et al (2008) who found that staff work backwards from an holistic judgement of the work, which is then 'compared with the set of criteria and standards and appropriate standards are then awarded in each of the criteria, ensuring the fit is commensurate with the holistic level of achievement awarded" (p135). One of

Tan & Prosser's (2004) conceptions of grade descriptors found amongst tutors (generic descriptors) also sees criteria as a 'postscript to the assessment process' (p271), something used to check whether students should be allowed to progress and not useful at all for feedback.

Other empirical studies have also demonstrated variations in marking criteria. For example Woolf (2004) found evidence of language used differently and subjectively by markers. Nonetheless communities of practice do provide a level of objectivity (Shay 2005) as shown in Baume et al's (2004) study of portfolio marking. Their examination of different staff marking decisions surfaced a number of shared criteria for judging student work.

The current study

These three critiques of the use of explicit and predetermined criteria and standards in higher education assessment form the backdrop for this study. They suggest that the assumptions on which our policy environment (and our staff development?) rests are fragile, not firmly supported in either theory or practice. Overall, the work, in particular of Sadler and Yorke, is revealing that "collectively, conceptions of standards of achievement may be less secure than many would prefer" (Yorke 2009: 72). Not surprisingly, there is a call for further research (Grainger et al, 2008; Sadler 2009a) but what studies exist, except for some honourable exceptions, tend to rest at the level of theoretical exposition in relation to how staff make marking decisions.

This is not surprising. As Orr and Blythman argue (2005), assessment practices are so 'naturalised' that it is hard to access them and thus the moderation meeting is 'a rare window into the multiple factors which influence academics' interpretive acts' (2005 :668) in a way that interviews may fail to do. Orrell (2008) identified the disjunction between tutors' espoused and actual practices when it comes to marking and therefore reinforces the view that conventional qualitative research, predominantly interviewing staff about their grading decisions, will not generate the type of understanding we are seeking.

This project, then, was designed to investigate how tutors go about the marking process, what strategies they use to arrive at a grade and how that is mediated through the use of artifacts such as written assessment criteria and standards. In exploring this topic, we purposefully selected research methods, in particular *think aloud protocols*, aimed at revealing practice, as opposed to espoused approaches to marking. Overall, the project sought to illuminate the gap between the widespread and largely unchallenged policy development in this field and the emerging critique set out above.

Using think aloud to investigate marking practice

A sample of twelve lecturers in a range of subject disciplines and professional fields were encouraged to think aloud as they graded two of their first year undergraduate students' written assignments. The think aloud activity was followed by a semi-structured interview that gathered data about their experience of grading student work and on the process of marking the two specific assignments. In addition, after each data collection, the interviewer recorded field notes concerning their perception of the data collection event. The sample included six academic tutors from each of two post 1992 universities in England. The two universities have considerable numbers of students on professional programmes including initial teacher education degrees and postgraduate studies.

Six of the tutors were in the professional field of teacher education and the other six were studying a range of other subject disciplines: history (2), English literature (2), business studies (1) and performing arts (1).

Think aloud protocols (Ericsson & Simon, 1993), recording participants as they attempt to verbalise their thinking during completion of a task, have been widely used to investigate problem solving and critical reasoning, mostly from an information processing perspective that attempts to build cognitive models of problem-solving strategies. Much of this work has focused on problem-solving by professionals working in health settings (eg. Ritter, 2002) but in the higher education context some has involved study of critical thinking by students (Phillips & Bond, 2004). The work has considered the focus of attention, what do problem solvers pay attention to, as well as cognitive strategies, what methods do problem-solvers use and how might these be modelled. Evaluation of think aloud method has considered to what extent the activity of thinking aloud might affect performance. In an example using recognition of analogies between narrative texts Lane & Schooler (2004) found by using a control group that thinking aloud appeared to impair recognition of deep analogies between stories and caused participants to focus on surface characteristics. In the area of assessment useful work using think aloud has been focused on school level external examiner marking practice (Suto, Crisp & Greatorex, 2008) and this has included some level of critical review of the data collection and analysis approach (Greatorex, 2008; Greatorex & Nadas, 2009). In these studies, at least from an information processing perspective and in simulated assessment contexts, the use of think aloud was found not to significantly affect the grades awarded.

Many studies use think aloud combined with another data collection instrument such as a semi-structured interview to gain a different perspective on the problem-solving activity. For example Orell investigated experienced academics comparing their beliefs about assessment, gained from interviews, with insight into their actual assessment practice gained through use of think aloud when marking scripts (2008). The current study follows this approach to data collection, tutors were audio recorded thinking aloud as they marked two student assignments and then audio recorded during a semi-structured interview.

From a socio-cultural theoretical perspective analysis of think aloud protocols may be useful in gaining insight into problem-solving but the assumptions underpinning the method differ from those applied by researchers working within a cognitive information processing framework (Smagorinsky, 1998; Ericsson, 1998). One of the key challenges of analysing think aloud from a socio cultural perspective is that the activity setting is as important as the think aloud protocol itself. In the current study this means that the marking activity needs to be seen as *situated* and that to understand the practice of the tutors the wider context and history of the activity need to be at least inferred. As Smagorinsky points out ‘interpreting a protocol requires knowledge of the participant’s cultural history, the researcher’s goal-directed behaviour within the conduct of the study, and the degree to which their congruence allows for words-as-signs to be assigned similar meanings by the two of them’ (1998: 165). From this perspective the analysis of the think aloud protocol must include some consideration of the response of the participant to the marking as a research data gathering activity. The use of a semi-structured interview after the think aloud activity was intended to provide some access to the wider contextual influences affecting the tutor during their marking activity. The researcher also wrote up field notes following each data collection event to provide some insight into the conduct and social interaction between participant and researcher. A socio-cultural perspective means that the words used by the tutor during the think aloud activity are seen as part

of a dialogue within a social context. The study relies on inference from the think aloud, the interview and the field note data to consider the wider historical and social context of the marking activity and the question of addressivity. The tutors may appear to be most immediately *addressing* the researcher and the research team but also other characters in the wider context including possibly the tutor's peers as second-markers and moderators, the students, the external examiner, the exam board, the wider subject discipline community including in the case of lecturers in professional fields the relevant professional bodies and practitioner community.

The think aloud protocols were analysed using a thematic qualitative analysis across the sample. An initial coding framework was constructed based on our reading of the literature in relation to the research design but this was developed in an iterative way through discussion within the research team as saturation in the data and early coding identified emergent themes and dimensions within themes (Ritchie & Lewis, 2003). The themes were developed further and all of the think aloud protocol data was coded through a constant comparative approach and memo writing to reach an established framework of conceptual themes and an initial understanding of the relationships between them.

In a second stage of the analysis each tutor in the sample was considered in turn and the understanding of their think aloud protocol, as indicated by the thematic analysis, was compared with an analysis of their interview transcript and the field notes describing the data collection event from the perspective of the researcher. The interview schedule began with an opening question 'what kind of marker do you think you are?' before focusing on the marking of the two specific scripts during the think aloud and the use of artefacts such as assessment criteria, marking schemes or grade descriptors within that process. Additional prompts then focused attention onto influences on the tutors approach to grading. The final question in the interview schedule asked tutors to reflect on their experience of participating in the think aloud process to provide a participant perspective on the influence of the data collection method. The purpose of the interview was to provide some insight into the perceptions of the tutors about their marking practice within their wider workplace context. The analysis of the transcripts involved iterative development of an initial coding framework based on the schedule and on our focus on interpreting the contextual influences expressed within the rules by which tutors are governed in their marking activity.

Tutor Think Aloud Transcripts

This section sets out the key categories emerging from the thematic analysis of think aloud transcripts, these are the transcribed recordings of the tutors attempting to verbalise their thinking by speaking aloud as they marked two student assignments.

Marking Focus

Through analysis of the think aloud transcripts the focus of attention of the tutors was identified by developing grounded themes arising from the data. The key categories emerging were entitled 'surface', 'global' and 'support' these are presented here.

When engaging with a student script all of the tutors generally appear to concurrently apprehend surface and global features but they varied in their explicit mention of the other four key categories

and in the pattern of their marking strategy. These variations are considered in the next section on marking strategy and in the later section setting out the 'profile' of each tutor as a marker.

The category of 'surface' was applied to tutors' think aloud comments focusing on apparently practical and relatively small scale issues that the student has or has not done correctly including spelling, punctuation, grammar and citation as well as use of side headings and formatting.

The category of 'global' was used in contrast to surface to include those comments in which the tutor appeared to be focusing on more holistic characteristics of the student script. For some tutors this was a focus on meeting the brief, has the student completed the task as set by the assessment guidance? Within this category other tutors appeared to use a conceptual artefact such as 'critical analysis' to judge the student script in a holistic way, if a script displays critical analysis then it is at least a 'pass'.

The category of 'support' was developed to include the tutors comments, positive and sometimes negative, when the student provided, or failed to provide, support for their developing argument. Tutors valued different kinds of support and there were signs of subject differences although the small sample size means that these can merely be noted tentatively. Many tutors commented positively when students provided support from the literature and were critical when they felt the students had made unsupported assertions. Some tutors, particularly the lecturers in teacher education, valued support from the practice of the student, where they referred to experience in their work based learning setting. Some tutors were keen for their students to provide concrete examples to illustrate their more abstract ideas and arguments.

Marking Strategy

Some categories were developed from the thematic analysis of the think aloud transcripts in relation to the marking process. These were entitled 'criteria', 'feedback', 'norm-referencing' ,'banding' and 'grade decision'. This section will introduce these categories and then go on to present the two broad patterns or strategies used by the tutors to mark a script.

The category 'criteria' was used to include all explicit comments by tutors in relation to the wide range of guidance, assessment criteria, grade descriptors and marking schemes that they might refer to during the marking process. This approach was found to be necessary because of the variation in terminology used by different tutors even within the same university. The explicit mention of criteria varied considerably between tutors. In the case of lecturers in teacher education some of the tutors pointed out that they were also assessing students against the professional standards for teachers or at least seeking evidence of professional values within the student scripts. One lecturer talked about review by the government agency Ofsted and what the inspectors would be looking for in student work and tutor feedback.

The term 'feedback' was used to categorise the tutor comments around what to include in their feedback to the student. Although feedback was particularly concentrated in the later stages of marking a script many tutors start to identify feedback very early in the process of engaging with a script and this provided some indication of their purpose in marking as being assessment for learning and their audience as being the student rather than other stakeholders in the assessment system.

The category of 'norm-referencing' was used when tutors explicitly used comparison with other scripts to help them in judging the student's work. Some tutors referred to norm-referencing in relation to the 'other' student script marked during the think aloud activity but overall the explicit use of norm-referencing was not very frequent, the social context of the data collection may have been influential in raising or suppressing its explicit use.

The terms 'banding' and 'grade decision' were simply used to categorise tutor think aloud comments in relation to locating the student work into a broad grading band or then assigning a specific percentage mark to the script.

Several tutors begin their scrutiny of a student script by considering the list of references provided, the bibliography, as a way of gaining an overview of the focus of the work and the level of scholarship involved. In general this appeared to be a strategy to consider the global characteristic of the script but even so some tutors would also comment on surface aspects of the citation and references at this point such as formatting.

Two patterns of marking strategy are apparent when the behaviour of each tutor is considered in relation to the thematic analysis. The first pattern involved an initial engagement with the student script with surface and global features commented on and in some cases explicit reference to the criteria. This initial appraisal leads to explicit 'banding' so that the script is placed in a grade category such as first class, upper second and so on. After the banding in a second stage the tutor continues to consider the work in a period of what appears to be checking, providing a rationale for the banding, as well as refining the banding to finally reach a grade decision that allocates a specific percentage mark to the script. The second pattern involves a similar initial pattern of noticing surface and global features but there is no second stage and the tutor moves through banding to a grade decision in rapid succession. A simple diagram of these two and one stage marking strategies is provided in figure 1. Eight tutors followed the same pattern for both of their student scripts, three used the two stage banding strategy and five used the single stage pattern. Three tutors used the two stage strategy for their first script but switched to the single stage for the second script.

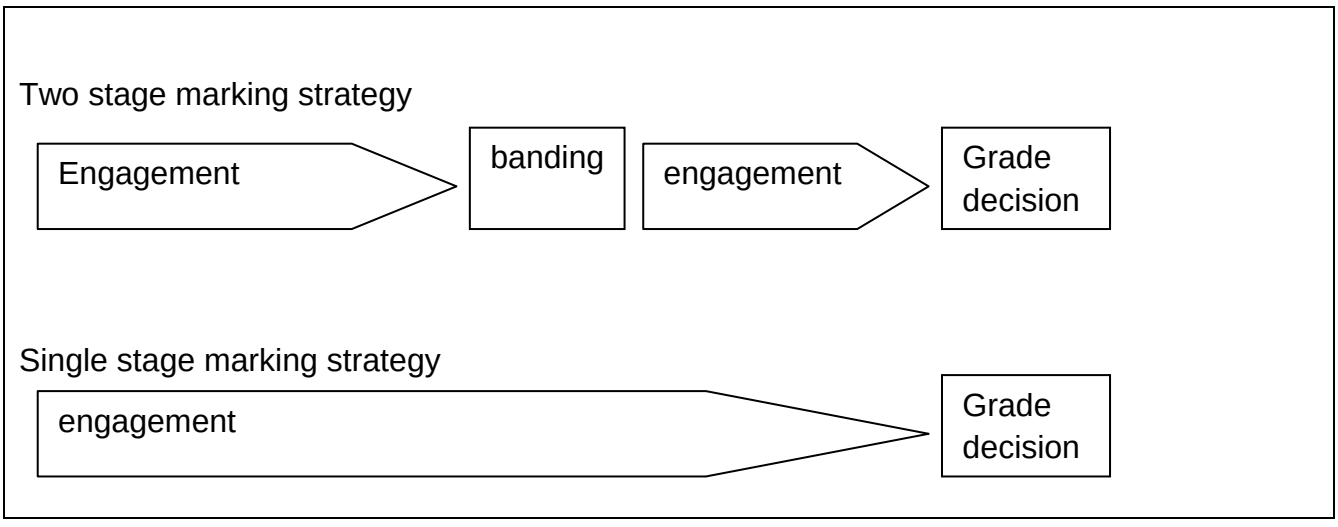

Figure 1: marking strategies used by tutors

The sample of tutors is too small to draw significant conclusions but there are some indications of possible subject discipline differences in the sense that only lecturers in teacher education used the two stage banding strategy. Even very experienced lecturers in teacher education, with 20 years in higher education, used this two stage banding strategy. Unfortunately the sample of tutors is not balanced in terms of experience and all but one of the teacher education lecturers had less than 10 years in higher education whilst the tutors in other subject disciplines all had 10 years or more. In terms of gender again the sample is too small to draw conclusions and the two marking strategies were used by both male and female tutors.

Tutor Interviews

This section outlines the key categories emerging from the thematic analysis of the interviews with tutors that followed the marking and think aloud activity.

Tutors identified 'time' as the key constraint on their marking practice. This was more complicated than simply having the time to complete the marking of scripts, they described the need to find blocks of time for marking, the need to be away from distractions, the lack of institutional and even line manager awareness of the workload involved in marking and also the pace of marking, the need to take breaks in order to keep fresh and maintain a fair approach for each student. They also described the lack of time for thorough second marking and moderation which they valued highly as professional learning opportunities.

Tutors valued second marking, moderation and teaching team activity involving blind marking as very effective professional learning opportunities. They complained that often these activities were constrained by lack of time. They claimed that double marking of dissertations was important because of the individual nature and significance of such student work and that this was again a useful professional learning process. The lecturers in teacher education valued their prior experience in assessment in schools, even marking of work by very young children, and in their think aloud and interview responses need to be considered in the light of the well-established emphasis in many schools on assessment for learning including criteria based assessment and formative feedback.

Although tutors valued professional learning through second marking and moderation activity, in general the tutors expressed confidence in the grades that they award, at least to the extent that student work would normally be at least within the correct band and that most errors in grading would be of little significance to overall degree classifications for students.

Overall there are signs of tension between the value that tutors place on professional learning in relation to marking, the broad confidence that they express in terms of the grades awarded to students, and the lack of time for marking and for learning from each other through second marking and moderation activity.

Tutor Profiles

As a second stage in the analysis each tutor was considered in turn and a profile of them as a marker was constructed. This included their marking strategy as identified through the thematic analysis but also their perspective of themselves as a marker, their engagement with criteria and other artefacts during marking and the focus of their attention during marking. Unfortunately Tutor 7, a lecturer in teacher education, completed the think aloud but not the interview and so it was not possible to construct a full profile for them. Whilst it is interesting to consider and compare the tutor profiles, including identifying patterns in relation to subject discipline and experience in higher education, it is important to note the limitations caused by the small sample size.

Four of the lecturers see themselves as 'fair' or at least average markers including three teacher education lecturers. Two position themselves as hard markers and just one sees himself as 'generous'.

Several of the lecturers admit to not using the concrete hard copy artefacts such as learning outcomes, criteria or grade descriptors and claim to have internalised these. In one institution the feedback sheet includes the list of learning outcomes and this clearly puts this right in front of the tutor when marking or at least when writing feedback, otherwise the lack of use of physical artefacts when marking may have been even lower. Experienced lecturers did not explicitly refer to 'criteria' including physical artefacts such as grade descriptors or marking schemes, except in the case of experienced lecturers in teacher education who did claim to use criteria including having physical artefacts present when marking.

Only three of the tutors, all teacher education lecturers, referred explicitly to 'criteria' in their think aloud protocol. Only teacher education lecturers used the two stage marking strategy with all 5 of them using it on at least one of their student scripts. Lecturers in other subject disciplines all used consistently the one stage strategy. All of the teacher education lecturers included explicit comments on feedback to the student in their think aloud whilst only half of the lecturers in other subjects mentioned this. There are then some tentative signs of subject discipline differences between teacher education lecturers and others.

The Think Aloud Process

At the end of the interview the tutors were asked about their perception of the think aloud activity. Half of the tutors found it interesting and useful in provoking their thinking about their marking. Just two of the lecturers felt that they had significantly affected their practice by making them more carefully articulate their judgements, 'I don't like to admit to being more rigorous. I think maybe I was'. Two admitted to feeling self-conscious at least during the marking of the first script and two confessed that it had probably suppressed their normal level of sarcasm or cursing about the students' work. Overall the tutors felt reasonably comfortable with the think aloud activity and did not see it as having an excessive impact on their practice.

The researchers conducting the think aloud recordings and interviews made brief field notes after each data collection event and these suggest that all but one of the tutors seemed fairly relaxed about the process of exposing their marking practice in this way. One of the tutors did however seem somewhat defensive about their practice of marking and sought to justify their approach to the researcher. As they had volunteered to take part it was not surprising that the tutors were all very

interested in the issues raised by the research project and they also were very interested in the think aloud method and expressed an interest in the findings of the analysis.

From the research team perspective the think aloud method appears to have produced rich data and our initial analysis has suggested some revealing patterns within lecturer grading practice. The semi structured interviews appear to need some revision because they did not reveal contextual influences on the lecturers as much as we had hoped. The interviews were left very open ended to avoid undue steering of responses but they did not for example provide much insight into how the accountability agenda in higher education contexts was shaping the thinking or behaviour of the lecturers.

Discussion

The analysis indicates that the explicit use by lecturers of criteria, including the text based artifacts of university assessment, was generally low although it does seem to be used as a check in the case of grades on a boundary or after an initial grade has been identified and this is in line with previous studies (Shay, 2004; Orell, 2008). Some teacher educators, perhaps because of their formal knowledge and practitioner experience in schools do appear to use, or attempt to use, criteria more explicitly.

The use of norm referenced assessment appears to be apparent in the marking practice of many of the lecturers although the nature of the data collection event may have contributed to comparison between the two student scripts. The use of norm referencing is officially frowned upon within the paradigm of criteria based assessment but the current study supports the arguments made by Sadler (2005; 2009a; 2009b) in questioning the idea that lecturers are able to use criteria in a process of analytical marking. It also challenges the way that students are currently sold the idea of criteria based marking as a way of reassuring them that the marking process is fair.

The study suggests that lecturers pay considerable attention to surface features of scripts during the marking process but also that they seek global features or characterstics, some of them also focus on identifying useful feedback for the student. The way that lecturers resort to the published criteria as a check on their grades suggests that they are not overly influenced by surface features in reaching their grading decision but this process deserves further attention. The significance placed by markers on surface features such as correct referencing was investigated by Grainger et al (2008). Most importantly the students may get mixed messages if lecturer feedback highlights surface features even if the grade awarded is more closely related to global characteristics.

The two different strategies for marking identified in this study do show some relation to Sadler's argument for the value of holistic approaches to marking (2009a; 2009b). The one stage strategy used by most of the lecturers especially those not in teacher education does appear to be close to a holistic approach provided not too much attention is given to surface features. The use of think aloud does appear to have provided some additional insight into the marking strategies of lecturers and this additional detail adds some value to the concept of holistic marking. This small pilot study also identified possible differences between subject disciplines that deserve further investigation.

The lecturers did express broad confidence in the marks they are awarding but the issue of 'time' for marking itself and for thorough collaborative work with colleagues in second marking and

moderation was raised as a key challenge. In developing assessment processes universities might do well to acknowledge the importance of this social element of agreeing standards and support it and resource it more formally by saving expenses currently allocated to other quality assurance activities. In an investigation of how academics engaged with assessment practice in their departments Jawitz (2009) argues that their workplace learning requires harmonization of their individual habitus with the collective habitus of the departmental community of practice. The current study supports the idea that useful academic development work needs to view the workplace learning of academics as a key element of ensuring fair marking and maintenance of standards.

The think aloud process was not seen as too intrusive by the lecturers themselves and observations by the research team also supported the idea that lecturers were generally not too defensive about this scrutiny of their practice. The steps taken in this study to adopt an ethical approach included using first year assignments and ensuring that those scripts involved were also second marked. The proposed study design was approved by the institution's ethical clearance committee procedures and overall the research team feel reasonably comfortable with the outcomes despite uncertainty about the impact on tutor grading and the possible repercussions for students. However this is a contentious area and many researchers and lecturers may have reservations about investigation of 'live' marking. Further study of 'live' marking of student work will need to consider carefully how much impact the think aloud process may have had on the grades awarded to the scripts involved. More focus on the script itself would help to pursue the subject discipline element of marking but may then draw in the student as a participant and require their permission. Perhaps in that case a study using scripts after formal grades have been awarded would be a more feasible approach and although it would lose the immediacy of using think aloud it would reduce the ethical concerns.

Conclusion

This small scale study has shown the potential of the think aloud method to give some insight into the mysteries of university marking. It provides tentative support for the critique of maintaining the current myth of marking as criteria based technical rational process and for the development of a more holistic approach that questions the socio-cultural context of university lecturers as they assess student work.

Using the interviews to investigate the context of the lecturers' marking practice was partially successful in this pilot study because it gave some insight into their perspectives on marking and the context of this work. However in the view of the research team this area of the study was weak and requires further strengthening in the larger scale project that we are planning.

The implications of the study for academic development work with departments and teaching teams include the need to acknowledge the place of norm referencing in the marking process, to critically consider marking strategies including final checking against the criteria and to recognise the importance of second marking, moderation and related discussion within the teaching team on workplace learning and quality assurance.

References

Baume, D., Yorke, M. & Coffey, M. 2004, "What is happening when we assess, and how can we use our understanding of this to improve assessment?", *Assessment and Evaluation in Higher Education,* 29(4), 451-477.

Biggs, J. & Tang, C. (2007). *Teaching for Quality Learning at University* (3rd Ed.), Maidenhead: Open University Press.

Blackler, F. (1995). Knowledge, Knowledge Work and Organizations: An Overview and Interpretation. *Organization Studies*, 6, 1021-1046.

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

Bloxham, S. & Boyd, P. (2007). *Developing effective assessment in higher education: a practical guide.* Maidenhead: Open University Press.

Bourdieu, P. (1988). *Homo academicus.* Stanford CA: Stanford University Press.

Bourdieu, P. & Wacquant, L. (1992). *An invitation to reflexive sociology.* Cambridge: Polity Press

Delandshere, G. (2001). Implicit theories, unexamined assumptions and the status quo of educational assessment. *Assessment in Education*. 8(2), 113-133.

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), 301-313.

Eisner, E.W. (1985), *The art of educational evaluation: a personal view.* London: Falmer.

Elton, L. & Johnston, B. (2002), *Assessment in Universities: A critical review of research*. York: Higher Education Academy.

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

Ericsson, K.A. & Simon, H.A. (1998). How to study thinking in everyday life: contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture and Activity,* 5(3), 178-186.

Engestrom, Y. (1987). *Learning by Expanding: an activity-theoretical approach to developmental research*. Helsinki: Orienta-Konsultit.

Engestrom, Y. (2001). Expansive Learning at Work: toward an activity theoretical reconceptualization. *Journal of Education and Work.* 14(1), 133–156.

Gonzalez Arnal, S. & Burwood, S. (2003). Tacit knowledge and public accounts. *Journal of Philosophy of Education,* 37(3), 377-391.

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

Greatorex, J. & Nadas, R. (2009) Using 'think-aloud' to investigate judgements about A-Level standards: does verbalising thoughts result in different decisions? *Research Matters*. 7, 8-16.

Greatorex, J. & Suto, W.M.I. (2008) What do GCE examiners think of 'thinking aloud'? Findings from an exploratory study. *Educational Research*, 50 (4): 319-331.

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

Knight, P. (2006). The Local Practices of Assessment. *Assessment & Evaluation in Higher Education,* 31(4), 435-452.

Knight, P.T. & Yorke, M. (2003). *Assessment, Learning and Employability. Maidenhead:* Open University Press.

Knight, P. & Trowler, P.R. (2001). *Departmental Leadership in Higher Education.* Buckingham: Society for Research in Higher Education / Open University Press.

Lane, S.M. & Schooler, J. W. (2004) Skimming the Surface: verbal overshadowing of analogical retrieval. *Psychological Science*. 15(11), 715-719.

Neil, D.T. & Wadley, D.A. (1999). A generic framework for criterion-referenced assessment of undergraduate essays. *Journal of Geography in Higher Education,* 23(3), 303-319.

O'Donovan, B., Price, M. & Rust, C. (2008). Developing Student Understanding of Assessment standards: a nested hierarchy of approaches. *Teaching in Higher Education,* 13(2), 205-217.

Orr, S. (2008, July). 'We kind of try to merge our own experience with the objectivity of the criteria': the role of connoisseurship and tacit practice in undergraduate fine art assessment. P*aper presented at the Assessment in Higher Education Conference, University of Cumbria, Carlisle*.

Orr, S. (2007). Assessment moderation: constructing the marks and constructing the students. *Assessment & Evaluation in Higher Education,* 32(6), 645-656.

Orr, S. & Blythman, M. (2005). Transparent Opacity: assessment in the inclusive academy. In C. Rust (Ed.) *Improving Student Learning: diversity and Inclusivity*. Oxford: OCSLD

Orrell, J. (2008) Congruence and disjunctions between academics' thinking when assessing and their beliefs about assessment practice. In C. Rust (Ed.) *Improving Student Learning: Theory Research and Scholarship.* Oxford: OCSLD.

Phillips, V. & Bond, C. (2004) Undergraduates's experiences of critical thinking. *Higher Education Research & Development*. 23(3), 277-294.

Price, M. (2005). Assessment Standards: The Role of Communities of Practice and the Scholarship of Assessment. *Assessment and Evaluation in Higher Education,* 30(3), 215-230.

Price, M. & Rust, C. (1999). The Experience of Introducing a Common Criteria Assessment Grid Across an Academic Department. *Quality in Higher Education,* 5(2), 133-144.

Quality Assurance Authority (QAA) (2006). *Code of Practice for the assurance of academic quality and standards in higher education, Section 6: Assessment*. Retrieved July 2, 2009 from http://www.qaa.ac.uk/academicinfrastructure/CodeofPractice/default.asp

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

Ritchie, J. & Lewis, J. (Eds.) (2003). *Qualitative Research Practice: a guide for social science students and researchers.* London: Sage.

Ritter, B.J. (2002) An analysis of expert nurse practitioners' diagnostic reasoning. *Journal of the American Academy of Nurse Practitioners*. 15(2), 137-141.

Sadler, D.R. (2009a). Indeterminancy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education,* 34(2), 159-179.

Sadler, D.R. (2009b). Transforming holistic assessment and grading into a vehicle for complex learning. In G. Joughin (Ed.) *Assessment, learning and judgement in higher education*. Springer.

Sadler, D.R. (2008, June). *How the use of preset criteria short-changes the students*. Presentation at Assessment Seminar, University of Edinburgh.

Sadler, D.R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education,* 30(2), 175-194.

Shay, S.B. (2004) The Assessment of Complex Performance: A Socially Situated Interpretive Act. Harvard Educational Review, 74(3), 307-329.

Shay, S. (2005). The Assessment of Complex Tasks: A Double Reading. Studies in Higher Education, 30(6), 663-679.

Smith, E. & Coombe, K. (2006) Quality and qualms in the marking of university assignments by sessional staff: an exploratory study. *Higher Education,* 51(1), 45-69.

Stowell, M. (2004). Equity, justice and standards: assessment decision making in higher education. *Assessment and Evaluation in Higher Education,* 29(4), 495-510.

Smagorinsky, P. (1998)Thinking and Speech and protocol Analysis. *Mind, Culture and Activity*. 5(3), 157-177.

Suto, I., Crisp, V. & Greatorex, J. (2008) Investigating the judgemental marking process: an overview of our recent research. *Research Matters*. 5, 6-9.

Suto, W.M.I. & Greatorex, J. (2008) What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, 34 92): 213-233.

Tan, K.H.K. & Prosser, M. (2004). Qualitatively different ways of differentiating student achievement: a phenomenographic study of academics' conceptions of grade descriptors. *Assessment & Evaluation in Higher Education,* 29(3), 267-281.

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), 479-493.

Yorke, Bridges & Woolf (2000) Mark distributions and marking practices in UK higher education: some challenging issues. *Active Learning in Higher Education*, 1 (1): 7-27.

Yorke, M. (2009). Faulty signals? Inadequacies of grading systems and a possible response. In G. Joughin *(Ed.), Assessment, learning and judgement in higher education*. *****

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

This document was added to the Education-line collection on 20 January 2010