Mark my words: the role of assessment criteria in UK higher education grading practices

Sue Bloxham, Peter Boyd and Susan Orr

Studies in Higher Education 36 (6) 2011, 655–670 DOI: 10.1080/03075071003777716

This article seeks to illuminate the gap between UK policy and practice in relation to the use of criteria for allocating grades. It critiques criterion-referenced grading from three perspectives. Twelve lecturers from two universities were asked to 'think aloud' as they graded two written assignments. The study found that assessors made holistic rather than analytical judgements. A high proportion of the tutors did not make use of written criteria in their marking and, where they were used, it was largely a post hoc process in refining, checking or justifying a holistic decision. Norm referencing was also found to be an important part of the grading process despite published criteria. The authors develop the notion of tutors' standards frameworks, influenced by students' work, and providing the interpretive lens used to decide grades. The implications for standards, and for students, of presenting the grading process as analytical and objective are discussed.

Introduction

Marking is important. The grades we give students and the decisions we make about whether they pass or fail coursework and examinations are at the heart of our academic standards. Assessment in higher education involves decentralised, subjectspecific decision-making processes, given credence in the UK by processes of quality assurance involving national agencies and review systems, external examining and local moderation. This quality assurance emphasis, whilst varying in detail across other English-speaking higher education systems, is currently located within a paradigm of accountability: explicit learning outcomes, constructive alignment (Orrell 2003), transparency and criteria-based marking (Quality Assurance Agency [QAA] 2006). As Grainger, Purnell, and Zipf (2008) point out, this pressure for accountability requires assessment decisions to be justified, and this is reflected in the QAA Code of Practice which specifies that 'Institutions [should] have transparent and fair mechanisms for marking and moderation' (QAA 2006, 16).

These 'validating practices' (Shay 2004) are designed to make the processes and judgements of assessment more transparent to staff and students, and to reduce the arbitrariness of staff decisions (Sadler 2009a). They assume that if the criteria are explicitly stated, they become accessible to all parties, albeit with a little help to understand the language. For example, the QAA code of practice states that institutions should 'make information and guidance on assessment clear, accurate and accessible to all … thereby minimizing the potential for inconsistency of marking practice or perceived lack of fairness' (QAA 2006, 8).

One result is that the 'production, publication and discussion of clear assessment criteria … [is now regarded as] 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, and there is some evidence that they can make a difference to student learning (Bloxham and Boyd 2007).

It has to be said that there is some confusion between criteria and standards (Grainger, Purnell, and Zipf 2008), and they are probably used interchangeably by different people and in different contexts. Sadler (2005) recognises them both as part of a criterion-based approach to assessment. Building on his four models of criterionreferenced assessment, he distinguishes criteria that are designed to judge how well the student has demonstrated progress towards the desired learning outcomes from standards which involve specifying qualitative criteria or attributes. In this analysis, standards can be seen in the typical department or university grade descriptors, which specify what students must do in relation to generic criteria 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. Whether stated as criteria or standards, they are a key 'validating practice' of universities (Shay 2004, 309); 'the mechanisms that academic communities put in place to ensure the validity of their assessment of student performance'. As Gonzalez Arnal and Burwood (2003) argue, the advocates for these tools believe that they help assure the quality of programmes by making explicit what is involved; a 'process of "exteriorisation" [which] makes judgements publicly grounded and thus objective' (380).

Sadly, given the resources deployed on these validating practices, they are founded on argument and assertion for the most part rather than empirical enquiry. Indeed, Sadler (2008) points out that the drive for the setting and use of assessment criteria does not have the theoretical or research support that one might have assumed. A survey of the literature indicates three key sources of criticism of this paradigm of accountability in relation to criterion-referenced assessment and marking: sociocultural, cognitive and empirical.

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. As Delandshere notes, drawing on the work of Bourdieu (1989), assessment is 'primarily understood as technology' (114), and this has disguised its role in social reproduction:

> The system of beliefs, values and purposes in which the agents involved are participating is rarely discussed. The perspectives taken when stating evaluative judgements are often assumed to be understood and agreed upon, when in fact they are rarely explicit or public, and hence, not open for scrutiny or discussion. (Delandshere 2001, 121)

Delandshere notes that assessment practice is based on assumptions that 'knowledge is monolithic, static and universal' (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 actually 'a context-dependent, experience-based and situational judgement' (323). However, a techno-rational view of knowledge underpins the accountability paradigm. It equates publication with explicitness and, as Gonzalez Arnal and Burwood (2003, 382) argue, this does not stand up to scrutiny. 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. (Original emphasis)

Other researchers also challenge the notion that it is possible to make explicit the tacit knowledge involved in assessment decisions (O'Donovan, Price, and Rust 2008; Orr 2007; Sadler 2009a; Shay 2005). 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 judgement which accrues through immersion in a subject discipline. This is an 'interpretivist' view of assessment which recognises the power of the local context (Elton and Johnston 2002; Knight and 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' (498).

However Shay (2005) asserts 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. Nevertheless, it allows for an element of professional and local interpretation, and there is considerable 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 (Ecclestone 2001; Knight and Yorke 2003; Price and Rust 1999).

Cognitive 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 and 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, and lastly that we are able to mentally manipulate a complex set of explicit criteria whilst reading student work in order to make a grading decision (Sadler [2009a] refers to this as analytic grading).

The 'accountability paradigm', as expressed in most guidance to tutors, advocates 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 and Wadley 1999; Orr 2008; Yorke 2009). In particular, Yorke makes the point that assessors' grading behaviour is tacitly influenced by norm referencing, and Shay (2004) and Orrell (2008) also found that tutors draw on their knowledge of different students' work in order to make their judgements. A further pressure on marking consistency is the open and diverse nature of student work. Shay (2004) argues that reliability has been based on highly standardised assessment tasks, whereas higher education assessment is characterised by low levels of standardisation. These complex tasks create problems for inter-marker reliability.

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. (177)

In other words, he is challenging the view that lecturers assess criterion by criterion, and arguing that they decide on the respective contribution of different criteria after they have made a holistic assessment of the work. He draws on research into complex judgements to argue that criteria often merge in practice or interact with each other, and this complex, 'fuller' process of judgement generates a 'truer' representation of the quality of the work than an analytic approach. Furthermore, staff use different sets of criteria for the same type of assignment (e.g. an essay). In addition, he makes the point 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, Sadler urges a greater acknowledgement of the contribution of holistic judgement.

Sadler's view that predetermined criteria do not reflect the full range of criteria being used to judge students' work is echoed in the work of Yorke (2009) and Tan and Prosser (2004). The latter found that some assessors consider that grade indicators do not 'depict the full range of desired qualities of student work' (273). However, Tan and Prosser's work found great variation in staff ideas about the power of grade descriptors to guide student expectations and staff assessment practices.

Empirical critique

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, that 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 do not feel confident getting on with making judgements. Over time they acquire the practical mastery which guides their decisions, and which is a form of socialisation that emerges from doing the job, marking over and over again. Indeed, assessors learn to mark by marking. Consequently Shay found that assessors were drawing on this 'feel for the game' (Bourdieu and Wacquant 1992) rather than published criteria when marking. Shay's respondents also identified the practical issues associated with trying to use written criteria, struggling with the 'false compartmentalisation' (2004, 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 a holistic approach to making judgements.

Orr and Blythman (2005) and Hawe (2003) 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 by tutors 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' (Orrell 2003, 198). Grainger, Purnell, and Zipf (2008) found that staff work backwards from a holistic judgement, awarding commensurate marks to individual criteria afterwards. One of the conceptions of grade descriptors found amongst tutors by Tan and Prosser (2004) also sees criteria as a 'postscript to the assessment process' (271).

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 academics do provide a level of objectivity (Shay 2005), as shown in Baume, Yorke, and Coffey's (2004) study of portfolio marking. Their examination of different staff marking decisions surfaced a number of shared criteria for judging student work. Research by Jawitz (2009) also indicates that staff gradually absorb the implicit assessment criteria of their departments, but this was not as a result of agreed criteria. He argues that processes such as double marking and feedback from the external examiner gradually harmonise the individual habitus with the collective habitus of the department.

Framing the current study

These three critiques of the use of predetermined criteria in university assessment form the backdrop for this study. They suggest that the assumptions on which our policy environment 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). The quality of existing research on grading is criticised (Orrell 2008) and what studies exist, apart from some honourable exceptions, tend to rest at the level of theoretical exposition in relation to how staff make marking decisions. Not surprisingly, there is a call for further research (Grainger, Purnell, and Zipf 2008; Sadler 2009a)

This project, then, sought to illuminate the gap between the widespread and largely unchallenged policy development in this field and the emerging critique set out above. It investigated 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 artefacts such as written assessment criteria and standards.

Using think aloud to investigate marking practice

Orr and Blythman (2005) argue that assessment practices are so 'naturalised' that it is hard to access them, and Orrell (2008) identified the disjunction between tutors' espoused and actual practices when it comes to marking. Therefore, we purposefully selected research methods aimed at revealing actual, as opposed to espoused, marking practices.

A sample of 12 lecturers were encouraged to think aloud as they graded two of their students' written assignments, which were either essays or similar discursive writing. The think aloud activity was followed by a short semi-structured interview that gathered some information on their experience of grading student work, and on the process of marking the two specific assignments, including the use of artefacts. In addition, after each data collection, the interviewer recorded field notes concerning their perception of the event. The sample was from 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 came from a range of other subject disciplines: history (2), English literature (2), business studies and performing arts. The tutors were volunteers, recruited through open advertisement in both universities.

Think aloud protocols (Ericsson and 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 informationprocessing 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 (for example, Ritter 2002), but in the higher education context some has involved study of critical thinking by students (Phillips and Bond 2004). The work has considered the focus of attention, what problem-solvers pay attention to, as well as cognitive strategies, what methods problem-solvers use and how might these be modelled.

Evaluation of the think aloud method has considered how it might affect performance. In an example using recognition of analogies between narrative texts, Lane and 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, research using think aloud has been focused on school-level external examiner marking practice (Suto, Crisp, and Greatorex 2008), and this has included some level of critical review of the data collection and analysis (Greatorex and Nadas 2009; Greatorex and Suto 2008). In these studies, at least from an information-processing perspective and in simulated assessment contexts, the use of think aloud was claimed 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 problemsolving activity. For example, Orrell (2008) 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. The current study follows this approach to data collection: tutors were audiorecorded thinking aloud as they marked two student assignments, and then audiorecorded 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 (Ericsson and Simon 1998; Smagorinsky 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 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 wordsas-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 context of the marking and this question of addressivity (Bakhtin 1986). 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 examination board, and the wider subject discipline or professional 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, 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 and 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.

For the purposes of this article, the findings have been developed only from the think aloud transcripts, supported by the field notes in relation to use of artefacts such as assessment criteria. In other words, we have attempted to explore the judgement processes staff appear to use rather than their espoused approach, with a particular focus on the role of 'criteria' in marking. The term 'criteria' is used to include all explicit comments by tutors in relation to the wide range of guidance, assessment criteria, grade descriptors, statements of standards 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.

Findings

The think aloud transcripts reinforce theoretical ideas about the complexity of the marking process in higher education. In almost every case, tutors appeared to come to a holistic conclusion regarding the final mark. They did not show evidence of linear or discrete processing of individual criteria;

indeed they appeared to be involved in multilayered juggling of overlapping elements in order to reduce them to a single representation in a percentage mark or grade. There was no evidence of them assigning marks to individual elements and then combining them to arrive at a final grade.

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 and noticing of 'cues' (Orrell 2008) in the student script, 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 cues in the script, but there is no second stage, and the tutor moves through banding to a grade decision in rapid succession.

In many cases, tutors made an initial judgement, and then referred to assessment artefacts such as grade descriptors to help them refine this grade to a specific mark.

OK. Now I step back from the essay and try and get an overall perspective on it. I've been thinking all the way through that it was a 2:1 and now I'm wondering if there's a possibility that it's a First. So I'm going to the Faculty of Arts assessment matrix which we are all supposed to use and which I find helpful as a rule of thumb. (T7)

OK. He concludes it quite well. So I'd say that's a good essay and I'm thinking it might be, it's certainly in the 60s. It might be a 70 so I'm just going to check. I've got a grid with the criteria here for the different marks that I might give. (T10)

This suggests that markers are not cynically referring to criteria post hoc in order to defend their judgements, but are using them to help refine 'hunch' decisions. This finding supports the post-judgement use of criteria in complex decision making discussed in the literature review. There was also some evidence that tutors used assessment criteria to help them turn a holistic grade decision into a justified grade decision; in other words, they appeared to have decided the grade before referring to the criteria, but sought support from the published criteria/standards to defend their decision.

Double-check. Coherent structure, clear writing, generally thorough. Some lapses in attention to detail. Referencing, professional standard with some lapses. Yeah it's a C but it's a high C so I'll give it 58 which means that all the assessed learning outcomes are satisfactory. (T8)

Although seven of the 12 tutors did make some verbal reference to criteria whilst marking, actual written criteria mostly seemed to be used in this post hoc way. We should probably not be concerned about this for two reasons. Firstly, following the discussion above, it may be the only way we can use criteria in complex decision making. Secondly, it retains an important role for published criteria in the judgement processes, particularly in negotiating exact grades to award and, therefore, some potential for inter-marker consistency.

Where staff referred to criteria during, rather than at the end, of marking a script, they appeared to adopt a 'threshold' rather than a 'standards' approach; that is, they were checking that the relevant element had been included rather than the standard of the work:

Then she goes on to say why she chose the Vikings – because of its significance, its importance in understanding what it is to be British and where it fits into the standard Scheme of Work. All those are things in the criteria so again I'll put a double tick in the margin just to remind me that I've ticked them off the criteria in my head as I do it. (T3)

We've now moved on to the analysis of it so she's covered the first point of the assessment criteria well. There's a clear statement and explanation of key concept and analysis. (T12)

Whilst this evidence indicates that assessment criteria are being used by many staff in their decision-making process, albeit in different ways, it begs the question, 'is this what students think is taking place?' The way criteria are presented to many students (if they consider them at all), they may assume some form of analytical marking is taking place, where their final grade is decided by weighting the contribution of individual criteria. Does assessment guidance lead students into thinking mistakenly that criteria are something fixed, accessible and rational that they can use to guide their work? Students' perceptions of marking processes is not something which has received much attention in research studies to date, although Carless's (2006) work on student perceptions of feedback did reveal concern that markers were not using the same standards, were not using published criteria or were using other criteria, such as perceptions of students' effort.

Use of physical artefacts

Drawing only on the think aloud data and the field notes, four of the 12 tutors did not have any assessment artefacts to hand whilst marking, and two further tutors had artefacts available but did not use them to make grading decisions. However, half the tutors did physically use artefacts, and it is probably not a coincidence that five of these six markers were teacher education lecturers. Indeed the sixth teacher educator did 'tick off' learning 'objectives', but did not explicitly use an artefact to arrive at a grade. It was a teacher educator who was the only tutor to look over the criteria before commencing marking. It could be argued that the teacher educators' use of artefacts reflects their 'assessment literacy' and prior experience of assessment in schools, where there is a well-established emphasis on criteria-based assessment. Alternatively, the inspection and audit regimes extant in UK teacher education may be the cause, and this deserves further investigation.

It is of significance, though, that all but one of the non-teacher educators did not use physical artefacts to support their marking. Despite the extensive critique of predetermined assessment criteria for marking, considerable effort is currently invested in creating them, and we frequently give students the message that they are a key aspect of our assessment procedures. Whilst the sample in this study is small and biased towards the humanities and social sciences, the findings do suggest that explicit use of assessment artefacts whilst marking may not be widespread in some academic disciplines. Whether this reflects individualised views regarding appropriate standards is less clear, and it is important to note that lecturers' apparently subjective assessment approaches might be better understood as shared and co-constituted. Shay uses the term inter-subjectivity to stress the social nature of lecturers' personal interpretive frameworks for judging assessment (Shay 2005). Adopting a theoretical perspective that draws on Bourdieu (1996), Shay argues that lecturers' assessment frameworks are socially produced. Indeed, there was no sense amongst the (non-artefact) markers in the present study that they did not have clear, legitimate standards against which they were making their judgements:

Thinking about the marking and reviewing it briefly in my head before I make any comments and just deciding into which ballpark area it fits. Is it the first, upper second, lower second, third, fail – it's not a fail because it does some of the things it says on the tin but on the other hand it's not a scholarly essay from a Year 2 student. It's something which is satisfactory and it does provide a rationale and it is quite practical but that's as far as it goes so it's probably in the 40s rather than in the 50s and that's probably what I think. Upper 40s rather than the lower 40s but I'm still thinking about that. (T3)

The essay is well written in the sense that there were no very obvious grammatical errors or spelling errors. It isn't you know written in a fantastic literary style you might want to reward but it is free of the errors that are common in probably 50% of the essays I'm marking. It's also well presented. It meets the criteria we expect and the bibliography especially is set out quite well. (T5)

I'm choosing the word 'good' there quite carefully. I use a series of words, I think, which clue me in to how I'm thinking. Satisfactory means broadly in the 40s; sound comes somewhere in the 50s. If it's good it's – in terms of that particular bit and if it's excellent then it's a sort of it's a very high mark. (T3)

Analysis of the interview data from this study, yet to be published, will explore whether this confidence is rooted in a perception amongst 'non-artefact' markers that they 'know' or have internalised the criteria and do not need to refer to them whilst marking.

There is limited explicit evidence of staff purposefully going outside the published criteria in making their marking decisions:

The next paragraph is talking about levels of history which is not in the criteria but … would get her some credit in my head as I read it. (T3) And challenging the criteria as set out:

Yes. Evaluation and reflection – I have difficulty with this one because it says for a 2:1, B, 60–69, all you need is 'straightforward broad evaluation and reflection' and my own view is that they should be able to do considerably more than that for a 2:1. (T7)

The findings of this study, therefore, tentatively support earlier research which suggests that staff ignore criteria, choose not to adopt them or use implicit standards which may not match those published to students. However, we are not arguing that this is necessarily a threat to standards, rather a reasonable response to the acknowledged difficulty of working with predetermined criteria and statements of standards. Nevertheless, the finding prompts awkward questions about the messages we give to students regarding marking decisions. A more honest approach would be to help students to understand that application of assessment criteria in higher education is a complex task involving professional judgement rather than measurement (Bloxham 2009).

Norm referencing

The term ‘norm referencing’ indicates that the student’s grade is dependent on others in the cohort (Orr 2008). The think aloud transcripts indicated that two-thirds of the markers used some form of norm referencing in making their marking decisions, either by comparing the two assignments they marked for the research or by referring to a wider group of assignments. It must be recognised that the research process, in asking tutors to mark two assignments, might have actively encouraged this approach. Nevertheless, there is significant evidence in the data to support the argument, discussed earlier, that the distinction between criterion and norm referencing is unhelpful.

Tutors explicitly used comparison to help them make their final grading decision. For example, one tutor consistently mentions the poor quality of the second assignment in comparison with the first. At the end, he decided that the mark he had awarded for the first was too low given the grade awarded to the second, and manipulated the marks to create greater differentiation. Previous marking also helped situate the quality of the work:

This is quite interesting because I’ve already read another student’s work who’s working on the same area, so in my head I’ve got something to compare it to. (T12)

I’m inclined to go with 56-ish, mid 50s. A bit more than that and I’ll compare that to another one later. (T9, second assignment)

Referring to other work helped test grade decisions and markers’ understanding of the strengths and weaknesses of work:

I’d like to get another one done so that I prove mainly to myself that that wasn’t too – I wasn’t too hard on it and that wasn’t a disaster. (T9, first assignment)

You know those Ds that I marked must have taken me an hour each even though they’re only 1000, because I was trying to think ‘what have they done?’ and it wasn’t until I got a good one that I realised what I think has been going wrong with them. (T6)

Comparing with the difficulties faced by other students softened critical judgements:

Just referring to them by title is not sufficient but I do know they all have huge difficulty referencing government documents because I’ve got lots of emails. (T6)

This reliance on norm referencing in marking is an important if unsurprising finding. Policies and practices to secure academic standards continue to rest on a fairly ‘objectivist’ rationality (Shay 2004), such as that presumed by criterion referencing against published standards. However, studies (including this one) are increasingly revealing that ‘assessments are quite often fuzzy measures relating to fuzzy constructs’ (Yorke 2008, 172). One explanation of the continued influence of norm referencing in making assessment decisions is that staff are unable to interpret semantically ‘loose’ criteria without some kind of personal ‘standards framework’ (Ashworth, Bloxham, and Pearce 2010). Jawitz’s work on how academics learn to assess reinforces this idea that relevant knowledge ‘can only be understood with the “interpretive support” provided by participation in the community of practice itself’ (2009, 603). Such a framework for judgement is constructed and reconstructed over time by reference to other students’ work, other tutors’ marking and moderators’ and examiners’ feedback. Whilst this framework is not whimsical (Sadler 1987), but clearly informed by their university and subject community, there is a suggestion in these findings that each new set of marking involves some refinement of a tutor's framework in the light of the specific task set and the scripts in hand. Indeed, it might be argued that those tutors in the study who did not explicitly norm reference were still drawing on personal standards frameworks developed partially through norm referencing, although perhaps not in relation to the current set of assignments (Orr 2008).

From this perspective, written criteria/standards only take on meaning once the staff apply their personal 'standards framework' to them, and other students' work is crucial in forming that framework. In addition, the fluid nature of these frameworks, apparently constructed and reconstructed as they are during the act of marking, further undermines the notion that we can predetermine and publish accurate criteria to students. As Orrell (2008, 259) points out, 'the qualities of other students' performances … do not provide a stable basis for maintaining standards because as a basis for grade decisions they are unpredictable and highly variable'. In that sense, this research strongly supports the critique of assessment criteria and marking asserted by researchers such as Sadler (2005, 2009a, b).

Anonymous marking

Anonymous marking is often requested by students as important in achieving biasfree marking by staff. Consequently, it was interesting to see the influence of knowledge of the student in staff marking processes. The research drew on staff from two universities, one with an anonymous marking policy (eight staff) and one where it was only required for examinations (four staff). Consequently, a third of the tutors knew whose work they were marking. In practice there was almost no reference to knowledge of the student in marking processes, although it must be recognised that tutors may have been more circumspect about this matter knowing they were being recorded. The sole quotation below indicates that prior knowledge of the student might influence the content of feedback:

I would say that she needs to proof read carefully though there weren't many mistakes and having known the student for four years there are great improvements in her writing ability. (T12)

Another tutor, working with anonymised work, indicated that tutors have expectations of students' achievement levels which might subconsciously influence marking:

At this point I can de-anonymise it to see who's the lucky recipient. It's … and that's the sort of mark I would expect from her. (T11)

Consequently, this study adds little to arguments for or against anonymous marking.

Surface features

In the analysis of think aloud transcripts, the category of 'surface' was applied to tutors' comments which focused on apparently technical and relatively minor tasks that the student had or had not done correctly, including spelling, punctuation, grammar and citation as well as presentation. There was significant evidence that staff place considerable focus on surface characteristics whilst marking, with regular comments on aspects relatively unrelated to the demonstration of learning, such as not putting quotations in italics, correct referencing style and missing apostrophes. Despite the frequent expression of these thoughts, it is difficult to detect whether such features are important in grading decisions and serve as major criteria for staff judgements. In general, the transcripts suggest that whilst staff consider surface characteristics to be important, they strive not make them significant in grading decisions.

Things that jump out at me aren't really very important, they're just trivial. Like the fact that they haven't used the tab function, they've just space-barred and therefore it looks really unprofessional but it's not really going to affect the marks that much. (T2)

The third thing she needs to think about is that the work is not always as grammatical as it might be. There are still proof reading errors and misspellings which create a bad impression. It hasn't been marked down to any significant extent because of that. (T3)

A focus on the transcripts, at the point of making a grading judgement, suggests that surface characteristics of work do not act as a major criterion. Indeed, 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. However, the focus staff place on them during marking, in their corrections and in their feedback may give students an inappropriate picture of their importance in achieving high grades.

Occasionally technical matters appeared to be central to a grading judgement:

The absence of a bibliography seriously limits the mark. This is a distinct shame as this is otherwise first class work. But it's first year so it's not exactly going to fail for that but [it's] an important marker. (T4)

I'm not going to penalise her too strongly for those sentences with the great long words in but it could have been, it could have been you know a high 60s mark. (T9)

… it's better than the previous one I was looking at but I can't give it a 2:1 mark because of the poor expression which is consistent. (T11)

Grainger, Purnell, and Zipf (2008) discuss staff predispositions and biases, for example towards the importance of referencing correctly. They suggest that these may be brought into play when the criteria applied do not easily 'fit' the work, although there was no evidence of that in this study.

Conclusion and recommendations

This study indicates that there is a disjunction between stated policies and actual practices in higher education marking, particularly in relation to analytical, criterionreferenced grading. This does not appear to be the result of a cynical rejection of accountability or a determined adherence to conventional practices. Indeed, the participating tutors, as self-selecting, appeared to be confident and conscientious markers. At one level, it could be argued that their behaviour is a rational response to the contradiction between espoused policies of accountability (that is, publishing and using assessment criteria) and the reality of professional judgement, with its tacit knowledge, complex and interrelated criteria and socially constructed 'standards frameworks'.

Whilst this situation may tend to contradict published guidance on how higher education marking is quality assured, we might argue that it is a ‘good enough’ (Elton and Johnson 2002) approach, where dependable outcomes emerge from sufficient consensus about what constitutes accepted knowledge, rules and procedures. Indeed, change of any significant kind is likely to be too time consuming to be practical. Nevertheless, the findings do suggest that we should consider how our quality assurance of assessment might better reflect the reality of marking practices.

Accountability might, for example, change a focus of scrutiny from demonstrating written guidance and standards into demonstrating clear efforts to build shared understanding of marking standards; talking more rather than writing more in an attempt to build and maintain consistent expectations. We might systematise the inevitability of norm referencing through this debate, and by agreeing exemplars of different standards. We might argue that academic departments should create the atmosphere where staff feel comfortable discussing their marking decisions, and this will only arise when tutors understand that marking standards are social constructions and not some fixed entity which they are more or less able to perceive. The research also suggests that, in training staff to mark, we should make them aware of holistic marking and reduce the anxieties prompted by attempting to create and combine marks for all the individual criteria. Furthermore, whilst helping students recognise the importance of avoiding plagiarism, staff should not overplay ‘surface’ features in feedback to students for fear of sending an inaccurate message about their importance in grading decisions.