Assessment and Evaluation in Higher Education Vol. X, No. X, Month 200X, pp. 000–000

Marking and moderation in the UK: false assumptions and wasted resources

Sue Bloxham [*]

[a]*University of Cumbria, UK;*

Abstract

This article challenges a number of assumptions underlying marking of student work in British universities. It argues that, in developing rigorous moderation procedures, we have created a huge burden for markers which adds little to accuracy and reliability but creates additional work for staff, constrains assessment choices and slows down feedback to students. In this under-researched area of higher education, the article will explore whether there are other ways to provide confidence in marking and grading. These might divert this energy into productive activities with useful outcomes for students and learning.

Keywords: Assessment; Marking; Moderation; External examining; degree classification

*Corresponding author. Centre for the Development of Learning and Teaching, University of Cumbria, Lancaster, LA1 3JD, UK.
Email: susan.bloxham@cumbria.ac.uk

Marking and moderation in the UK: false assumptions and wasted resources

Introduction

This article is designed to add to the debate that is taking place in the pages of this journal on the inherent frailty of marking practices and variability of standards (for example, Baume *et al.*, 2004; Norton, 2004; Price, 2005; Read *et al.* 2005; Sadler, 2005; Hartley *et al.* 2006; Knight, 2006). The general lack of discourse on marking in higher education allows assumptions of reliable standards to continue largely unchallenged perhaps because, as Price (2005) suggests, it is too uncomfortable to discuss these matters which are at the foundation of our awards. This article will examine some of the assumptions evident in the higher education community:

1. We can accurately and reliably give a mark to most students' work;
2. Even if individuals' marking may sometimes be inaccurate, internal moderation ensures fair and appropriate standards in marking;
3. Even if internal moderation does not reflect expected standards, external moderation ensures students are assessed against consistent standards across the UK University sector;
4. Students' final award (degree classification) reflects their achievement in a consistent way within and, to a certain extent, across universities.

There are additional parallel assumptions regarding the 'validity' of assessment tasks in all its various forms. Needless to say 'accuracy' of marking as used in the above assumptions has a clear link to, and may have an impact on validity. For example, where students perceive that the tutor is using marking criteria that do not reflect the

published guidance, they may direct their efforts to matters divert the task away from assessing the intended learning. However, this paper does not pursue the issue of validity in HE assessment, not because it is taken as a given, but because it is a large and challenging issue beyond the scope of this paper.

Whilst recent reports suggest the UK degree classification system is no longer fit for purpose (Universities UK, 2006; QAA, 2007), they continue to assume that, at one level, marking is relatively reliable and accurate; meaning that marking is dependable and reflects the expected standards of work at the given level of study. These reports may argue that classification procedures aren't consistent or that grades awarded differ by subject discipline but they don't challenge the assumption that the grades students are given in the first place are reliably or consistently awarded. Perhaps the principal reason for this rests in a general confidence in our marking and moderation practices. Students, for the main part, do accept the grades they are awarded even if they recognise that staff give credit for different qualities when marking (Crook *et al.*, 2006).

A whole gamut of procedures is designed to support this general confidence: assignment guidelines, assessment criteria, grade descriptors, marking schemes and evidence of moderation. External examiners are considered a key guarantee of this confidence and universities still place considerable credence on their ability to assure appropriate and reliable standards (Watson, 2006). The following sections will examine the above assumptions.

Assumption 1: We can accurately and reliably mark students work

Higher education assessment is internally set and marked and, whilst this offers a level of autonomy not enjoyed in other sectors of education, it means marking is largely in the hands of the subjective judgement of tutors and assumes that academics share common views regarding academic standards.

The process of marking in higher education has not been examined in depth (Yorke *et al.*, 2000; Smith & Coombe, 2006) and what research there is shows that staff vary considerably both in the marks they give and in the shape of their mark distributions (Heywood, 2000). Elton and Johnston (2002) describe marker reliability as low for essays and problem style examinations except where mere knowledge recall is required and Knight (2006:440) suggests that the ability to measure students' achievements reliably may be more possible in subjects such as the natural sciences. He contends that 'non-determinate' subjects that deal with the 'human world' such as the arts, humanities and social sciences rely more on the subjective judgement of assessors.

The causes of unreliability are related to the nature of what is being measured by assessment in higher education. Knight (2003) argues that complex learning cannot be reduced to something simple enough to measure reliably; the more complex the learning, the more we draw on 'connoisseurship' (Eisner, 1985) rather than measurement to make our judgments. Elton and Johnson (2002) provide an excellent discussion of the literature in relation to a key dilemma in higher education between assessment for 'certification' (that is providing the means to identify and discriminate between different levels of achievement) and assessment for learning; setting out the

different positions of the positivist and the interpretivist approaches. Essentially, a positivist approach believes in the importance of validity and reliability, assuming that objective standards can be set. The alternative, interpretivist, approach rejects objective truth and conceives of assessment as based on a local context, carried out through the judgment of experts in the field. In their view, it is a social practice whose credibility emerges from a community of practice which shares a consensus about what constitutes accepted knowledge, rules and procedures. It is a 'good enough' (p39) approach in which 'dependendability' is parallel to reliability in positivist assessment' (Elton & Johnston, 2002:46).

An interpretivist view would argue that there is a level of professional judgment in most elements of undergraduate assessment whatever the discipline, particularly if we take Knight's (2006) position that all graduates should be learning complex capabilities such as initiative, adaptability and critical thinking. According to Broad (2000) conferring grades to complex written work is impossible and misleading. It is interesting to see that the Quality Assurance Agency in the UK (2006) is advocating broader grades or mark bands perhaps recognising the difficulty of precise percentage grading.

Academics, as 'connoisseurs' are considered able to make expert and reliable judgments because of their education and socialisation into the standards of the discipline (Ecclestone, 2001). Knight (2006) argues that this situated and socially constructed nature of standards means that measurement of complex learning can only take place effectively within its context, a local judgement made within that social environment be it a course team, department or subject discipline. Thus it is not

surprising that many studies have found considerable discrepancy between tutors in their assessment criteria and the grades they accord to assignments (for example (Baume *et al.*, 2004, Norton *et al.*, 2004, Price, 2005). They are using locally constructed and tacit standards to make their decisions. Overall, there is 'an underlying softness in the data that typically goes unrecognised' (Sadler, 2005:182).

Varying professional knowledge, experience and values leads to staff attaching importance to different qualities in student work (Read *et al.*, 2005; Smith & Coombe, 2006). The research in this field remains very limited but efforts to increase the reliability and validity of marking such as assessment criteria, grade descriptors and marking schemes are somewhat undermined by the difficulty of communicating their meaning (Price & Rust, 1999; Ecclestone, 2001) and by tutors' customary approaches to marking. Wolf (1995) contends that markers acquire fixed habits in their marking which they may not be aware of but which can influence their grading, and assessors may not understand or agree with the outcomes they are supposed to be judging (Baume *et al.*, 2004). Evidence also suggests that staff ignore criteria or choose not to adopt them (Price & Rust, 1999; Ecclestone, 2001; Smith & Coombe, 2006) or use implicit standards which may contradict the official standards (Baume *et al.*, 2004; Price, 2005; Read *et al.*, 2005). Generic institutional or departmental standards are not seen as robust by staff, creating difficulty in applying them to a specific module (Price, 2005).

Despite this unreliability, experienced assessors come to see themselves as expert markers (Ecclestone, 2001). Their judgments become more intuitive than conscious, as they develop 'mental models' of marking which they apply regardless of marking

guidance (Ecclestone, 2001). In reality, studies have found that experienced markers are no better than novice markers at applying standards consistently partly because new markers pay greater attention to marking and marking guidance (Ecclestone, 2001; Price, 2005). However, casual staff, in particular, may feel under pressure to mark generously when they face evaluation by students and fear poor appraisal following low marks (Smith & Coombe, 2006). Overall, the intuitive and essentially private nature of marking (Ecclestone, 2001; Smith & Coombe, 2006) and a lack of assessment scholarship and discourse amongst academics (Price, 2005) are not helpful in addressing these issues.

The issue of variation in tutors' marking applies particularly to the interpretation of assessment criteria and marking standards. Woolf (2004) argues that criteria only make sense in context. They often include words such as 'appropriate', 'systematic' or 'sound' which are relatively meaningless unless you have a framework in which to understand them. We can try to write criteria and standards more explicitly but there will always be a degree of professional judgment which comes from being a connoisseur in the discipline. 'Even the most carefully drafted criteria have to be translated into concrete and situation-specific terms' (Knight & Yorke, 2003,:23). Therefore research (Price, 2005; Swann & Ecclestone, 1999a) suggests that technical changes to practice such as marking grids and assessment criteria are insufficient on their own because application of a marking scheme to a specific assignment is a 'social construct' negotiated between the members of that assessment community and influenced by their tacit knowledge (Baird *et al.*, 2004). On a more positive note, other researchers (Klenowski & Elwood, 2002) believe that common standards do

become established amongst cohesive staff teams, and this is certainly a view frequently declared by tutors.

Assumption 2: internal moderation ensures fair and appropriate standards in marking

Anxieties about standards of marking have contributed to a growth in procedures to assure standards, in particular the request that 'Institutions have transparent and fair mechanisms for marking and moderation' (Quality Assurance Agency, 2006:16). But do these help?

Moderation is a process for assuring that an assessment outcome is valid, fair and reliable and that marking criteria have been applied consistently. Interestingly, whilst reliability is discussed in depth, 'moderation' as a term hardly surfaces in higher education assessment literature. There are a number of benefits considered to accrue from effective moderation. These include improved reliability resulting from the opportunity to discuss differences in the interpretation of criteria and marking schemes, prevention of assessment being 'unduly influenced by the predilections of the marker' (Partington, 1994: 57) and militating against the influence of 'hard' or 'soft' markers. In addition, transparent moderation procedures are likely to increase students' confidence in marking and they provide 'safety in numbers' (Partington, 1994), giving staff confidence in dealing with students (Swann & Ecclestone, 1999a). Finally, seeing others' marking and discussing marking decisions can have an important role in staff development and the creation of an assessment community amongst marking teams (Swann & Ecclestone, 1999a).

Moderation is often in the form of second or double marking of summative assessment and, although published studies of its effectiveness are lacking, evidence suggests that staff see it as essential for providing fairness to students and assuring the quality of work (Hand & Clewes, 2000). According to Partington (1994) time-consuming double marking ceases to be necessary if there are published mark schemes moderated by external examiners. However, as discussed previously, marking schemes can be interpreted differently or even ignored. Partington (1994) discusses the difficulty in second marking with regard to convergence of the two markers when the second marker knows the mark which has been awarded by the first marker. He suggests this convergence is likely to be aggravated where the second marker is moderating the work of a more experienced colleague. Alternatively, when the mark is not known (blind second marking), the individual markers might be expected to value the characteristics of the students' work differently and ultimately this may lead to students being treated more fairly. On the other hand, Hornby (2003) discusses the concept of 'defensive marking'. Where an assignment is 'blind' double marked, a tutor may feel at greatest risk from internal or external moderators and may practise 'risk averse' marking, erring towards giving average grades.

One key difficulty in sample second marking is the dubious assumption that a sample can be 'taken as indicative of the whole' (Partington, 1994: 2). This is an important, and underdiscussed, issue. The assumption that a whole set of work has been accorded fair marks because a second tutor agrees with the marking standards applied to a 10% or 20% sample is erroneous. It assumes that each tutor marks all pieces consistently even though they are trying to apply multiple, complex, assessment

criteria. Again, moderation through sample second marking may be helpful where the criteria are focused on lower levels of learning, but it is easy to see that, at higher levels of HE study, sample moderation has its limitations in terms of ensuring fair and consistent standards in marking.

Assumption 3: external moderation ensures students are assessed against consistent standards across the sector

The external examiner system is designed to bring a level of external accountability to assessment decisions, that is to ensure that standards are comparable with similar awards elsewhere, and to ensure that the institution's academic regulations and assessment procedures are effective and fairly applied. External examiners carry out their role using a range of processes such as meeting with staff and students and reviewing course documents (Higher Education Academy, 2004). However, their most central task continues to be the moderation of examination scripts and coursework assignments to test standards and facilitate the comparability of treatment between students and with other institutions.

There is limited research evidence on the effectiveness of external examiners but what exists does not inspire confidence (for example, Silver *et al* 1995). Their impact is regarded as 'light touch' (Murphy, 2006: 40) and unreliable (Price, 2005) and the task is considered to have become very difficult as the nature of higher education programmes has become more complex and modularised (Heywood, 2000, McGhee, 2003). A recent report (QAA, 2007: 1) argues that the ability of external examiners to check whether students are being treated fairly in comparison with those in other

institutions may be limited by the extent of their experience. Nevertheless the external examiner system has been regarded by some as guaranteeing comparable quality in British higher education (Heywood, 2000), 'a figure of immense moral importance, significantly envied in other systems' (Watson, 2006:2).

Evidence (QAA, 2005) suggests that most UK institutions have now established sound external examining procedures in terms of appointment, induction, powers, communication and reporting but, as with other aspects of quality assurance (QA) in higher education, these assure the reliability of the procedure rather, perhaps, than the quality of the underlying practice.

Assumption 4: Students' final award (degree classification) reflects their achievement in a consistent way within and across universities.

Fortunately, at the level of degree classification in the UK, this issue is now firmly in the public domain with various reports (Universities UK, 2006; QAA, 2007) identifying that procedures for classifying degrees are inconsistent both within and between universities. Various uniform findings have emerged from the work of the Student Assessment and Classification Working Group regarding disciplinary and university differences in student grading (for example Yorke *et al* 2000). Bridges (1999) found evidence of significant discipline-related marking differences with 'qualitative' subjects such as history and English creating very different mark distributions from 'quantitative' disciplines such as Mathematics. It is not surprising, therefore, that researchers report significant differences in the proportions of 1st class honours graduates across disciplines ranging from 21.1% in Mathematics to 3.7% in

Law (*Yorke et al*, 2000). The evidence (Universities UK, 2004, 2006) is that current arrangements do not allow stakeholders to assume that degree classifications provide any level of comparability across or within institutions and some students are particularly disadvantaged by this, for example those on 'joint' programmes. The QAA report argues that students' achievements are affected by disciplinary differences in marking practices as well as institutional rules on how grades are combined to provide an overall classification. Whilst the QAA summary suggests that staff in the HE sector will understand what the differences mean, others will not be knowledgeable about disciplinary and university differences reflected in different rules and that something based more clearly on achievement would be more suitable.

What are implications of these false assumptions?

The foregoing discussion suggests that, at best, many key assumptions on which our marking and moderation practices are based are unverified, or at least, in need of much more research and development. The intention here is not to disparage this situation but to recognise it as a necessary corollary of the nature of the university enterprise. Undoubtedly, there are things we could do better with greater attention to assessment and marking, but the evidence suggests that subjectivity is unavoidable.

However, this does present us with a quandary. In essence, assessment provides the basis for assuring academic standards (Price, 2005). Institutional accountability in relation to assessment has been a high priority in recent years and clear procedures and rules facilitate the external scrutiny that has been demanded. Whilst much of this growth in quality assurance may not increase the impact of assessment on student

learning, it does allow the system to be 'judged in relation to its overall coherence and transparency' (Crook *et al.*, 2006: 96). However, although reliable procedures give the appearance of 'good order', they do not necessarily deliver good quality assessment practice (Crook *et al.*, 2006). It could be argued that institutional energy has focused on equitable and consistent assessment procedures at the expense of developing assessment practice. The gap between procedure and practice is reflected in a conspicuous divergence between how well institutions think they do assessment and general student dissatisfaction with it, for example in relation to the helpfulness of feedback (Hounsell *et al.*, 2006; National Student Survey, 2006; Crook *et al.*, 2006).

Indeed, such QA procedures may have a potentially detrimental effect on student learning, with the illusion of confidence created by such QA procedures skewing assessment design away from that which supports learning towards that which serves mainly 'certification' and 'quality assurance' (providing evidence to judge the appropriateness of standards on the programme (Gibbs, 1999)). For example, extensive internal moderation may delay the return of work to students despite the evidence that timely feedback is important for learning (Gibbs & Simpson, 2004-5). Procedures such as anonymous marking can contribute to a dislocation between author and reader in higher education assessment with impersonal feedback appearing irrelevant or inaccessible and lacking dialogic quality (Crook *et al.*, 2006). A further example is where markers are asked not to write comments on work as this may prejudice the double marker although such comments may feed forward into the students' future learning.

Biggs (2003) suggests that the use of external examiners may discourage innovative assessment practices as institutions restrict themselves to approaches that can easily be understood out of context. In this way assessment procedures designed to improve comparability of standards may conflict with the use of a range of assessment methods and limit the use of more innovative methods of assessment with their demonstrable benefits (Bloxham & Boyd, 2007). An example would be the rejection of peer assessment despite its role in helping students understand the standards required in their writing. Biggs (2003) suggests that external examiners cannot be fully aware of, and in sympathy with, the aims of the teaching and the approach to assessment of a programme and yet responses to their comments are part of the QA system; hence the pressure on tutors to use 'traditional', well-understood assessment methods.

In addition, the pressure for reliable marking can skew assignment choice as, for example, an over-riding concern for demonstrably reliable marking may prevent the use of group assignments and peer assessment or may encourage use of assessments that usually foster low level, determinable, learning such as multiple choice tests (Scouller & Prosser, 1994).

The picture presented above suggests that it would be foolish to propose simple solutions to some of the problems of marking and moderation. Indeed Knight considers that 'solutions are not to be had' (Knight, 2006: 450). It may well be argued that standards are sound despite weaknesses in practices and procedures; we have a 'good enough' approach which generally inspires confidence. However, an alternative view would be that assessment is not in good order. In today's mass higher education,

assessment can use more resources than teaching (Gibbs, 2006) including considerable resources devoted to moderation efforts which are largely unrelated to student learning and assessment receives lower scores for student satisfaction than any other factor (National Student Survey, 2006).

Potential changes

Can we learn from other sectors of education? Murphy (2006) suggests that university assessment practice lags well behind its equivalent in the school sector. Public examination boards and many professional bodies invest a considerable amount in designing, checking and moderating examinations such that levels of reliability and validity are high. The massive resource for this level of investment is generated by the large numbers of candidates taking individual courses which provide economies of scale not available in the HE sector.

It is easy to see why this approach does not easily transfer to higher education. Firstly, there is a debate whether the complex learning required in university assignments can be described in marking schemes that do not lay themselves open to substantial differences in interpretation and therefore the 'training' requirement for markers would be considerable. A second problem with this approach, one interestingly that GCSE examination boards are having to tackle, is that it drives you away from coursework towards controlled conditions such as examinations, because it is only in controlled conditions that you can assess both the process and the product. Assessment by coursework allows students to be involved in very different processes, yet come up with products of a similar standard (Knight, 2000). Finally, the additional

resources required to improve reliability and accuracy through this 'public examinations' approach would, seemingly, add little to student learning. Indeed unseen examinations have generally been found to encourage inferior learning (Gibbs & Simpson, 2004-5).

Taking the programme approach

A second approach would 'accept and embrace the subjectivity of judgment' (Clegg & Bryan, 2006: 224). It would place value in the professional judgment and experience of teaching staff but would not assume that there is a correct mark for each individual assignment or examination script and thus would not waste time using moderation for that purpose with the majority of assignments. The focus would shift from individual assessments to the overall profile of a student on the basis that a series of marks awarded over a period of time might provide a more accurate assessment of student (Heywood 2000). A review of the calculations involved in arriving at the classification of a degree in the UK makes a strong case for moderating at the programme level where a student's achievement is represented by a single grade or figure (a situation unlikely to change in the medium term, I suggest). The following worked example illustrates that even major changes in marks for individual assignments are unlikely to influence a student's classification.

Worked Example

Let us suppose that a student's classification is based on their mean module marks, and achievement is weighted 40%:60% between level 2 and level 3 (intermediate and honours level). Each student completes eight modules at each level and the

assessment for most of those modules comprises coursework (50%) and examination (50%). Therefore each assignment/ examination contributes 2.5% of the final average at level 2 and 3.75% at level 3. If a student has a mean of 57% (a 2ii degree), a mark of 52% in a level two module contributes 1.3 marks to the final mean and even if moderation changed the mark to 100%, the mean mark would only increase to 58.2%. Likewise, a level 3 mark of 68% would need to be raised to 135% in order to move their average up to 59.5% and, thus, a 2.i. degree. Alternatively, the rise in classification of degree could be achieved by getting an extra 67 marks across level 3 only (4.19 marks per item) or an extra 100 marks across level 2 only (6.25 marks per item) or an extra 2.5 marks per item across the whole of both levels.

Moderation at the individual assignment level rarely, if ever, makes changes of the order or in a consistent direction as discussed in the *worked example* and thus the impact of individual item moderation on a student's overall profile is likely to be very limited. Yet, modular curricula with large numbers of assignments tie up significant amounts of staff time in moderation procedures.

A programme approach assumes that although each assignment's mark may have limited reliability, confidence should come instead from the professional judgement of several different tutors across a large number of different assessment opportunities. This approach suggests that we stop accepting that there is a 'correct' mark for each assignment or examination paper and agree that the range of marks gives us a sufficiently consistent picture of student achievement with careful moderation reserved for those students whose pattern of marks isn't sufficiently consistent or is borderline.

Undoubtedly a certain amount of item-level moderation would still be useful: for failing work, for new types of assignments, for new markers and to help staff develop mutual understandings of assessment criteria. In the latter cases, a pre-marking discussion of a sample of assignments might satisfy those important tasks more effectively than second marking.

Module level approach

An alternative, and not uncommon, approach is to carry out moderation through examining mathematical differences in student achievement at the module level. It is relatively easy for two tutors to agree over the marking of an individual item but it is in the patterns of marking that we may begin to identify systematic differences in marks between different groups and different teachers and it is perhaps those that we should be paying more attention to. For example, investigation could focus on sets of work where individual tutors' means and standard deviations fall outside the norms of other modules completed by the same or similar cohorts of students within and outside the subject area.

Involving students as partners in assessment

Recent theoretical development in the field of feedback is focusing on the importance of student as self-assessor who, in addition to receiving the tutor's feedback, is able to provide their own feedback because they understand the standard they are aiming for and can judge and change their own performance in relation to that standard, that is

self-regulation (Nicol & Macfarlane-Dick, 2006). This is assessment *as* learning, and is firmly located in Sadler's (1989) view that improvement involves three key elements: students must know what the standard or goal is that they are trying to achieve, they should know how their current achievement compares to those goals and they must take action to reduce the gap between the first two. Thus, as Black and Wiliam (1998a: 15) assert, 'self assessment is a sine qua non for effective learning' and certainly systematic reviews of research (Black & Wiliam, 1998a; Falchikov, 2005) indicate strong positive results and benefits to students of being involved in their own assessment.

Theoretical explanations for this view lie in the notion that part of being a subject specialist is the capacity to assess quality in that field. Involving students in assessment provides an authentic opportunity for them to learn what 'quality' is in a given context – solving a problem, doing an experiment, creating a design, or writing an essay – and applying that judgement to their own work (Black *et al.*, 2003). Thereby the student becomes aware of what the goals or standards of the subject are (Earl, 2003), a precondition of taking responsibility for their work (Swann & Ecclestone, 1999a). This view is supported by Black et al (1998a) when they stress that self assessment is the key to learning from formative assessment. It is not enough for a tutor to tell a student what they need to do to improve if the student does not understand what these comments mean in relation to the subject or their writing. They cannot take action to do anything about it until they begin to share the tutor's conception of the subject (Sadler, 1989).

Therefore, rather than see students as the recipients of our judgment, however subjective, we should involve them in that judgment process. Just as a doctor will share with a patient uncertainty about a diagnosis, so we should help students to

understand that application of assessment criteria in higher education is a matter of professional judgment, not a matter of fact. In other words, should we be gradually inducting students into the subjective nature of marking, increasingly expecting them to demonstrate why they think they have met the criteria?

Conclusion

Perhaps the most immediate conclusion is that, when developing their moderation policies, institutions should think carefully about the effective use of staff time and resources in specifying requirements for sampling and the extent to which second marking is obligatory. The marking task needs to be a balance between maintaining standards and the confidence of external stakeholders, practicability, quality assurance and the promotion of student learning.

However, at heart this is an epistemological issue; how is the knowledge of what is a good exam answer, essay, project or piece created? It is created through a social process involving dialogue and experience and using artefacts such as assignment guidance and assessment criteria but, in essence, it remains essentially an individual construct, heavily influenced by traditions in the subject discipline. Staff who work closely together may develop shared understandings and, therefore, in the local setting, there is greater potential for reliable marking. Nevertheless, subjectivity and differences within and across universities remain a difficult, if largely uninvestigated, field where research is clearly overdue. In particular, more research is needed to explore patterns of marks in higher education and what, if any, impact moderation has upon them.

The brief discussion of possible ways forward poses more questions than it answers. Undoubtedly, we do need to ensure that key stake holders maintain their confidence in marking but we need to do that, not through focusing on unattainable reliability and accuracy, but by emphasising professional judgment moderated by others in the higher education community, both internal and external, and including students themselves.

Acknowledgment

The author wishes to thank Nigel Appleton for assistance with preparation of this manuscript.

REFERENCES

Baird, J., et.al. (2004) What makes marking reliable? Experiments with UK examinations, *Assessment in Education: Principles, Policy & Practice,* 11 (3), 331-348.

Baume, D., et.al. (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. (2003) *Teaching for Quality Learning at University.* 2nd edn. (Buckingham, Open University Press).

Black, P., et.al. (2003) *Assessment for Learning: putting it into practice.* 1st edn. (Maidenhead, Open University Press).

Black, P. & Wiliam, D. (1998a) Assessment and classroom learning, *Assessment in Education,* 5 (1), 7-74.

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

Bridges, P., & Bourdillon, P (1999) *Discipline related marking behaviour using percentages: a potential cause of inequity in assessment.* Assessment and Evaluation in Higher Education 24 (3), 285-301.

Broad, B. (2000) Pulling Your Hair Out: Crises of Standardization in Communal Writing Assessment, *Research in the Teaching of English*, 35 (2), 213-260.

Clegg, K. & Bryan, C. (2006) Reflections, rationales and realities, in: Bryan, C. & Clegg, K. (Eds) *Innovative Assessment in Higher Education.* (London: Routledge), 216-227.

Crook, C., et.al. (2006) Assessment Relationships in Higher Education: The Tension of Process and Practice, *British Educational Research Journal*, 32 (1), 95-114.

Earl, L.M. (2003) *Assessment as Learning.* (Thousand Oaks, California: Corwin Press).

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).

Falchikov, N. (2005) *Improving assessment through student involvement.* (London: RoutledgeFalmer).

Gibbs, G. (2006) Why assessment is changing, in: Bryan, C. & Clegg, K. (Eds) *Innovative Assessment in Higher Education.* (London: Routledge), 11-22.

Gibbs, G. (1999) Using Assessment Strategically to change the way students learn, in: Brown, S. & Glasner, A. *Assessment Matters in Higher Education: Choosing and Using Diverse Approaches.* (Buckingham: SRHE/ Open University Press), 41-53.

Gibbs, G. & Simpson, C. (2004-5) Conditions under which assessment supports student learning, *Learning and Teaching in Higher Education*, 1 (1), 3-31.

Hand, L. & Clewes, D. (2000) Marking the Difference: An Investigation of the Criteria Used for Assessing Undergraduate Dissertations in a Business School, *Assessment and Evaluation in Higher Education*, 25 (1), 5-21.

Hartley, J., et.al. (2006) What Price Presentation? The effects of typographic variables on essay grades, *Assessment & Evaluation in Higher Education*, 31 (5), 523-534.

Heywood, J. (2000) *Assessment in Higher Education.* (London: Jessica Kingsley).

Higher Education Academy. (2004) *Working paper 4: The process of academic external examining.* Available at:

http://www.heacademy.ac.uk/resources.asp?process=full_record&section=generic&id=366 accessed 20/11/2006.

Hornby, W. (2003) Assessing using grade-related criteria: a single currency for universities?, *Assessment and Evaluation in Higher Education,* 28 (4), 435-454.

Hounsell, D., et.al. (2006) *The Quality of Guidance and Feedback: paper presented to the Northumbria EARLI SIG Assessment Conference, 30th August-1st September.* Conference paper edn. Darlington:

Klenowski, V. & Elwood, J. (2002) Creating communities of shared practice: the challenges of assessment use in learning and teaching, *Assessment and Evaluation in Higher Education,* 27 (3), 243-256.

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

Knight, P.T. (2000) The Value of a Programme-wide Approach to Assessment, *Assessment and Evaluation in Higher Education,* 25 (3), 237-251.

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

McGhee, P. (2003) *The academic quality handbook : enhancing higher education in universities and further education colleges* (London: Kogan Page).

Murphy, R. (2006) Evaluating new priorities for assessment in higher education, in Bryan, C. & Clegg, K. *Innovative Assessment in Higher Education.* London: Routledge., 37-47.

National Student Survey. (2006) *The National Student Survey.* Available at: http://www.thestudentsurvey.com/. Accessed 13/9/2006.

Nicol, D. & Macfarlane-Dick, D. (2006) Formative assessment and self-regulated learning: a model and seven principles of good feedback practice, *Studies in Higher Education,* 31 (2), 199-218.

Norton, L. (2004) Using assessment criteria as learning criteria: a case study in psychology, *Assessment and Evaluation in Higher Education,* 29 (6), 687-702.

Norton, L., et.al. (2004) *Supporting diversity and inclusivity through writing workshops. Paper presented to the International Improving Student Learning Symposium, Birmingham, UK, 6–8th September.*

Partington, J. (1994) Double-marking students work, *Assessment and Evaluation in Higher Education,* 19 (1), 57-60.

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.

QAA (2007) *The Classification of Degree Awards*. (Gloucester: Quality Assurance Agency).

QAA (2005) *Outcomes from institutional audits: external examiners and their reports.* (Gloucester: Quality Assurance Agency).

Quality Assurance Agency. (2006c) *Code of Practice for the assurance of academic quality and standards in higher education. Section 6: Assessment of students.* 2nd edn. (Gloucester: Quality Assurance Agency).

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

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

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

Scouller, K.M. & Prosser, M. (1994) Students' experiences in studying for multiple choice question examinations, *Studies in Higher Education*, 19 (3), 267-279.

Silver, H., Stennett, A. & Williams, R. (1995) *The external examiner system: possible futures.* (London, Quality Support Centre).

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.

Swann, J. & Ecclestone, K. (1999a) Litigation and learning: tensions in improving university lecturers assessment practice, *Assessment in Education*, 6 (3), 357-375.

Universities UK (2004)*Measuring and Recording Student Achievement (The Burgess Report)* London: Universities UK

Universities UK. (2006) *The UK Honours Degree: Provision of Information.* (London: Universities UK), (2nd Consultation Paper)

Watson, D. (2006) *Who killed what in the quality wars.* (Gloucester: Quality Assurance Agency).

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, M., et.al. (2000) Mark distributions and marking practices in UK higher education, *Active Learning in Higher Education*, 1 (1), 7-27.