Red Guides
Paper 41

Assessment matters

Lesley Matthews

Red Guides address educational and staff development issues within Higher Education and are aimed at colleagues within the University and at other institutions. Some describe current good practice in Higher education, others evaluate and/or comment on curriculum development and many provide ideas for teaching. All are meant to stimulate discussion, initiate action and implement change.

The author (s) of this guide:
Lesley Matthews
School of Built Environment

Northumbria University

New edition: ISBN 978 1 86135 424 2

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Introduction

This guide is intended as an introduction to current debates and issues in good assessment practice in higher education. It does not claim in any sense to be comprehensive - which would require several volumes, given the rapid expansion of research into assessment issues over recent years - but rather attempts to introduce some of the key debates around both the purposes of assessment and aspects of good practice in assessment, as identified and explored in recent research. It raises questions as much as it supplies answers, with the intention of encouraging tutors to think more carefully and deeply about the purpose and nature of their assessment activities.

The purposes of assessment in higher education

Assessment matters. It is highly significant for both students and educational institutions. It is used to permit tutors to make judgements about standards achieved and to determine whether a student passes (and at what level of achievement) or fails. As such, it will influence not only whether students pass their course, and how well, but also the 'performance' achieved by the course overall (via 'progression' and pass rates) and hence, collectively, perceptions of the 'success' of the course and the institution. But perhaps most importantly for tutors and students, it matters because of the potential impact it can have on student learning. According to Trotter (2006), *"…researchers are of the opinion that one of the best ways of improving student learning is by altering student assessment"* (p505). This places assessment at the heart of effective learning, and indicates clearly why all tutors need to concern themselves with assessment issues. As Maclennan (2001) observes - drawing on the work of Crooks (1988) and Gibbs (1999) – "*the quality of student learning is as high (or as low) as the cognitive demand level of the assessment*" (p307). If the assessment requires deep learning, then students will find it necessary to engage with deep learning.

However, whilst the potential contribution of assessment to student learning is widely accepted in higher education, it should be recognised that its role in judging, and ranking, students' achievements is not without controversy. As Broadfoot (2000, p ix) points out, *"From its modest beginnings in the universities of the eighteenth century and the school systems of the nineteenth century, educational assessment has developed rapidly to become the unquestioned arbiter of value……….Equally remarkable has been the lack of any serious challenge to this hegemony".* It is possible, therefore, that we place too much faith in the validity of the outcomes of assessment. Leathwood (2005, p310) highlights what he sees as the political dimension to assessment, and argues that, *"assessment has served……to legitimise and rationalise the unequal distribution of power and resources in society".* In recent years, government policies have increasingly focused on the inequalities inherent in educational achievement in higher education, and specifically on the relative lack of participation of individuals from lower income groups. Pressure has increased, on the 'old' universities especially, to recruit more students from state comprehensive schools with historically low participation rates. Nevertheless, the notion of 'gifted' children remains inherent in policy approaches. As Leathwood (2005, p311) argues, there seems to be a generally-accepted view that the distribution of natural or inherited 'ability' in the population is in the shape of a pyramid, with a few of exceptional ability at the top, and a mass of those with low ability at the bottom. It is widely assumed that assessment should enable us to distinguish between these layers, and it is on this premise that much of the current debate on 'dumbing down' in higher education rests; for example, Chris Woodhead, former head of OfSTED, has called for 'intellectually rigourous academic education *which by* *definition only a few can benefit from'* (quoted in Leathwood, 2005, p311). As Leathwood points out (p313), "*a moral panic about lowering standards is an annual occurrence in the British tabloid press*".

In addition, the current emphasis on using the results of assessment – via percentage of first class degrees, failure and progression rates, etc - as a means to compare the 'performance' of different universities is also seen as contentious. It is argued that the extent to which the qualifications (such as degrees) which result from assessment in higher education can have universal and directly comparable meaning may, necessarily, be limited by virtue of the fact that assessment happens in specific and unique contexts (Knight, 2006). He argues that as "*assessment data are created in particular educational contexts……they are contexted judgements of contexted and complex achievements*" (p 435), hence each university's 'warrants' (awards) will differ for many reasons, including differing interpretations of the subject, different assessment criteria, varying mark ranges, and differences in mark aggregation practices.

In distinguishing between the 'learning' and 'judging' roles of assessment in higher education in the UK, tutors frequently refer to the 'formative' and 'summative' roles of assessment. More recently, the term 'assessment for learning' has become widely used to emphasise to tutors the potentially critical role of *all* assessment in developing student learning. Essentially, formative assessment will provide feedback from which students can learn how they might improve knowledge and skills, and hence future performance. It need not be a formal activity or even in written format, and much effective formative assessment occurs in seminar sessions and other classes in which students have the opportunity orally to 'test' their understanding and knowledge. The main purpose of formative assessment is, however, student learning.

Summative assessment – assessment that contributes a mark or grade which 'counts' towards an award – can, however, also have a formative role, particularly if it occurs at a point in the module which permits student learning to occur in time to influence subsequent assessment activity. Knight (2006) distinguishes between *feed-back* – which explains the judgement reached by the assessor, and is the type which is common in end-of-year assessments – and *feed-forward*, which is advice about how to improve future performance. Arguably, however, even 'feedback' on assessment should permit students to transfer these judgements to improve their future learning capacities. As McDowell et al (2006) argue assessment is an <u>essential</u> part of developing student learning, and "*formative assessment is central to effective teaching and, by engaging students in it as active participants, the effect is multiplied*" (pp 2-3).

Assessment also has aims which go beyond the immediate, educational context. If we hope to engender the capacity in all students to engage in independent, lifelong learning, then students need to be encouraged to undertake systematic self-assessment; this will help them to identify weaknesses and reflect on their own particular learning needs, which will assist them to develop as truly independent learners. Knight (2006) recognises that, outside of an educational context, we all routinely make informal 'assessments' - of situations, people, events, etc – and, in this wider context, he identifies *three* types of assessment purposes:

- *background assessment,* which is "*the judgements we make in the normal course of thinking and acting*" (p 442). He argues that, in education, tutors should aim to make this a 'foreground' activity by creating interactive tasks which involve plenty of opportunities for students both to make judgements and also to have discussions about those judgements (p441). This assists students to understand “the ‘rules of the game’ and what counts as acceptable, strong and indifferent achievement” (p441). To best facilitate this, it also requires that teachers concern themselves with the quality of the learning environment. If they succeed in ‘foregrounding’ the background assessment, then they have helped to create learning oriented assessment.

- *learning-oriented assessment*, which involves students and teaching staff using and applying criteria to evaluate their activities. Learning opportunities are provided through these ‘assessment conversations’ (p442), which enable students to gain experience of using criteria to make judgements about both their own and others’ work. Teaching staff need to provide both ‘feedback’ and ‘feed forward’, as this is crucial to assisting students to develop self-assessment skills and hence the capacity for lifelong, independent learning.
- *warranting achievement*, which is the ‘summative’, judgmental role of assessment. For students, this is usually (and understandably) perceived as the main purpose of educational assessment, as it generally results in a tutor-assigned mark or grade to indicate the ‘quality’ of the students’ achievements. However, tutors do need actively to encourage students to engage with the feedback and feed forward elements, as it is these which can positively influence students’ learning and future performance.

Recent trends in assessment practice

It is widely accepted that assessment practices in UK universities have been changing significantly over recent decades. A HEQC report in 1997 identified a gradual shift away from what it called the ‘traditional’ approach to assessment, which was based largely on examinations at the end of the academic year, and involving all staff in close scrutiny of questions and of marking schemes. Two trends - the expansion in student numbers and the modularisation of programmes - have arguably resulted in a greater variety of approaches to assessment, but possibly also in much looser connections between assessors. Holroyd (2000) identifies a number of ‘general patterns of change’, including:

- Greater emphasis on assessment for learning enhancement rather than for ‘certification’, and hence on formative rather than summative assessment
- Greater use of the ‘standards’ model of assessment; this involves criterion-referenced assessment (assessment against pre-determined criteria), rather than the ‘measurement’ model involving norm-referencing (meaning, assessment against group standards, involving ranking of individuals’ performances)
- Less dependence on a single method of assessment at the end of the course, changing to a variety of assessments within the course
- More involvement by others in assessment, such as self, peers, and workplace assessors.

In addition, the expansion of government involvement in standards in higher education, principally through the Quality Assurance Agency (QAA), and the emergence of *benchmark statements* to define *what* is to be assessed – though not *how* it is to be assessed – has undoubtedly exerted further influence on recent trends. According to the QAA (1996, p 13), assessment in modules should:

- Be in relation to outcomes made explicit to students, staff and employers

- Be based on a range of strategies through which a student can demonstrate what he or she knows, understands or can do
- Use a range of evidence
- Include review and reflection, to lead to the identification of future goals
- Facilitate formative recording of achievement
- Enable students to gain credits for their attainment.

A key concern in relation to the 'warranting' (summative) role of assessment must be to ensure that the possibility of variability in assessor judgements is limited by the assessment design. This is assisted by ensuring that assessments conform to consistent practices. For example, according to Northumbria University (2004 pp 4-6), good assessment practices should be:

(i) valid

This essentially means that it should assess what it claims to assess. It should match learning outcomes, be at the right level and accord with subject threshold standards.

(ii) reliable

Marks should reflect performance (suggesting that the full range may be appropriate), with appropriate criteria and relevant to that level.

(iii) consistent

Closely connected to reliability, this will be demonstrated by the involvement of other staff, with moderation of planned assessments, double/blind marking, use of external examiners, etc.

(iv) diverse

This means choosing the best methods to fit the topic/subject and the students' learning needs. It implies that, on most modular programmes, there should be a variety of assessment approaches, to reflect the diversity of the subjects studied and the desired learning outcomes.

(v) efficient

This has become ever more important as student numbers increase and staff workloads become ever more pressurised. It refers to the fact that assessment must be manageable within normal workloads, for both students and staff. Assessment must not demand excessive effort from either the assessed or the assessors.

(vi) understandable

This relates to the clarity of the tasks set, to ensure that '*the standards achieved are explicit and available*' (p 5)

(vii) support learning

Assessment should require students to engage with 'deep learning', learning by doing, encouraging the application of theory at appropriate levels.

(viii) provide effective feedback

This recognises the formative role of all (even summative) assessments, so feedback (in whatever form) should highlight how the work could be improved. It should also be timely; in fact, from the student's perspective, the sooner the better, but certainly no later than 4 weeks after submission (though particular Schools and subject groups may well have their own norms in relation to this). Some types of assessment lend themselves to very fast or even immediate feedback, such as computer-based tests, or oral presentations. It is also worth noting here that effective feedback is likely to occur frequently and often informally in many lessons – for example, when students are engaged in problem solving, debating issues, working in groups, working on designs, undertaking laboratory work, etc. This will often be oral feedback, although, possibly, many students will fail to recognise this as 'feedback' unless it is made explicit.

It remains best practice to consult with colleagues when devising assessments, as "*the benefit of having questions criticised by colleagues has long been accepted"* (Holroyd, 2000, p34). They can help to identify not only errors and ambiguities, but also may be able to suggest improvements which could strengthen, for example, the reliability or validity of the assessment. In many universities, it is common practice to have some form of formalised moderation process, often within a subject or discipline based team, but also commonly involving the external examiners, especially for examination questions (where students may have no opportunity to seek clarification of unclear questions).

Assessment methods

Summative assessment, for award purposes, needs to test the extent to which the student has successfully achieved the learning outcomes of the module or programme. However, as discussed above, it is clear that all assessment methods should facilitate learning and so have a formative role, by offering both feedback and feed-forward (Knight 2006). When determining the most appropriate assessment methods, a number of issues need to be considered:

a. what type of knowledge are we assessing?

Cranton (2006) argues that we need first to be clear about the type of knowledge we are assessing if we are to select appropriate assessment methods. She draws on the work of Habermas (1971) to identify three 'domains of knowledge', which, she argues, are acquired through differing research methodologies, and this understanding can be used to inform the assessment methods which are appropriate for different types of student learning. The domains of knowledge which she distinguishes, and their associated research methods (p 3), are:

a. *instrumental knowledge*, which is objective and empirically derived, such as understanding cause and effect relationships. This is acquired through methods such as quantitative measurement, experimental design and the traditional scientific method. To assess instrumental knowledge, we can use quantitative, objective strategies, as it is possible to identify if the student has supplied the 'right' answer. Hence, short answer and multiple choice tests, laboratory work, and computer-based exercises may all be appropriate here.

b. *practical knowledge*, which is about groups, relationships, culture and morality, so encompasses an understanding of our social and political systems. This is acquired through qualitative data, derived from methods such as interviews, texts and conversations. Since the student needs to be able to interpret their knowledge, assessment must provide opportunities for freedom of expression. Cranton suggests that essays, oral presentations, role plays etc are appropriate vehicles for assessing this, but, controversially, she argues that we are ill-advised to try to make 'objective, scoring judgements' about these 'interpretive procedures', by having pre-determined criteria. She asks,
"*what about the student who has the courage to challenge existing viewpoints*?", and argues strongly that *"good interpretive evaluations are trustworthy and credible…..we rely on the expertise, professionalism and credibility of the teacher…..talk of bias is irrelevant….."* (p6).

Holroyd (2000) asserts that assessment is "*a matter of judgement based on inferences from evidence evaluated as valid, reliable and fair*" (p37), and, similarly, Knight (2006) agrees that "*assessment is a practice of judgement*" (p 436). This viewpoint, which O'Donovan et al (2000, p74) refer to as the ‘connoisseur’ model of assessment (“most often likened to the skills of wine tasting or tea blending”), runs contrary to recent UK trends, which increasingly place emphasis on criterion-referenced assessment. Knight (2006) argues that only when the assessment is concerned with ‘observer-independent natural phenomena’ – which exist regardless of whether or not there are people – can we feel sure that the scope for assessor judgement is limited and hence the reliability of the ‘measurement’ (mark) determined by the assessment is most certain. He argues that “*some achievements lend themselves to quasi-assessments and many do not*” (p438), and Yorke et al (2000) confirm that “*some outcomes, such as the production of an artefact in art and design or the interpretation of a literary text, are not amenable to precise specification in advance and therefore have to be open to flexibility on the part of the assessor*” (p23).

Perhaps, then, the best that assessors can achieve in such cases is to try to ensure that, in designing assessments, they try to limit the possibility of wildly differing interpretations of the same evidence?

c. *emancipatory knowledge,* which is that acquired through critical reflection and self-reflection, so, to asses this, students must be involved in the self-evaluation of their work. Cranton argues that, as teachers, “we cannot say, ‘*Our goal is self-direction, self-determination and self-reflections, but I will judge how well you did’”* (p6). However, *“this is not to say that students can assign themselves marks or grades without dialogue and guidance……..”* (p5). As many experienced teachers have discovered when trying to introduce self- and peer-assessment into their evaluations, self assessment is not a skill which most students have acquired from their earlier studies. They need first to learn how to engage in self-evaluation; for example, with help to clarify learning goals and by seeking to validate their perceptions with the views of others.

b. What learning outcomes are to be assessed?

This point is closely linked to the previous observations about ‘type of knowledge’. Some learning outcomes lend themselves very well to particular forms of assessment rather than others. For example, if a learning outcome is ‘interpret information from a set of financial accounts’, then this lends itself well to being assessed via a fairly short, class-based test or examination, because there will be fairly brief and ‘right’ answers. On the other hand, if the learning outcome is ‘evaluate social policies to tackle multiple deprivation’, this may be best assessed through an extended piece of coursework, which permits the student better opportunities to demonstrate their application of evaluative skills. Practical, skill-based outcomes are best assessed through practical exercises or tests, such as undertaking presentations, laboratory experiments, completing a drawing, or building a model. The key issue here is *validity*: does this method of assessment best permit the student to demonstrate the achievement of the desired learning outcomes of this module (or element of the module)? As Yorke et al (2000) point out, “*intended outcome and assessment method have to be coherent*” (p 22). In this context, it is also important that assessments are sufficiently complex and challenging to permit a wide range of possible responses, so that the depth and breadth of student learning can genuinely be tested. Unless questions permit significant diversity in answers, there is a danger that tutors will need to rely on very minor differences to generate mark differentials.

c. Equality and fairness

Universities today have a much more diverse group of students than they once had, and this diversity brings with it additional challenges for tutors, to ensure that assessments do not advantage or disadvantage particular groups of students. For example, there is a widely held perception that some types of assessment favour one gender over another. This is summarised by Woodfield et al (2005) as, essentially, that girls have benefited from changes in modes of assessment, which explains the fact that they now generally outperform boys in academic achievement. They quote Pirie (2001), who claims both that exams have been 'feminised' and also replaced to some degree by continuous assessment. This view argues that coursework is better suited to the female approach to studying, while exams suit male study methods better. Martin (1997) attributes the lower proportions of female firsts at Oxford partly to greater exam-anxiety levels amongst women, but she also argues that their academic style is more cautious and less confident than men's (p480).

However, Woodfield et al's recent (2005) research at the University of Sussex found that women there outperform men on *both* forms of assessment, and that (contrary to expectations); male students as well as females expressed a preference for coursework over unseen exams. Also, as they point out, both genders perform better when coursework forms part of the assessment, and both genders felt that this provided a fairer, better measure of their educational achievement (p41). Gibbs and Lucas (1997) and Simonite (2003) both found that average marks were higher on modules with a higher proportion of coursework than on those relying more on traditional unseen exams, and Gibbs & Lucas argue that the trend of increasing reliance on coursework helps to explain the upward drift in degree classifications over recent years.

Madaus (1994) argues that, since all assessment ('testing') is underpinned by cultural values, and different cultural groups potentially have different intellectual traditions, minority ethnic groups may face problems with the values underpinning the approach to assessment. To guard against assessments favouring particular groups, Leathwood (2005) argues that *“lecturers need not only to be monitoring achievement and progression rates on their programmes, but also taking action to ensure that students from all groups are equally able to achieve the highest levels”* (p318). So, if it becomes apparent that particular groups of students have underperformed (relative to their other work) on a particular module, then this may suggest a need to review (for cultural or other bias) the assessment which was set.

d. Plagiarism and academic misconduct

As Woodfield at al (2005) point out, there are “*greater problems of author authentication issues associated with coursework against examinations*” (p46). While we can be reasonably confident (with identity checks and adequate invigilation) that an exam script is the student’s own work, this can not be said of most coursework. However, replacing all coursework with exams is unlikely to prove satisfactory, because arguably, it is more difficult to design traditional exams to test a number of higher-order skills, such as critical evaluation or the ability to synthesise complex information. Perhaps partly due to changes in the way that UK school awards (such as A levels) are assessed, students are also, arguably, less well-prepared or practised in the skills required successfully to take examinations than they once were.

To try to minimise opportunities for students simply to copy the earlier coursework of others, it is good practice to ensure that the focus and form of any assessment changes sufficiently each year. If possible, it can help to provide a very specific, novel scenario, within which the student has to apply their knowledge. It is also worth considering whether the assessment design could permit the process (of developing the answer) to be in some way be observed and/or assessed. Perhaps some of the work required by the assessment could be undertaken under supervision; for example, the staff supervision which occurs for a dissertation module will usually limit the possibilities for students simply to insert large sections of someone else's work.

To limit the potential for 'ghosting' of coursework (getting someone else to produce the answers, often 'to order'), assessment can be time-limited, with, say, a take-away paper which has to be completed and returned within 24 or 48 hours. Or, students could be supplied with a long list of potential questions at the outset, and advised that one will be selected later, for completion within a short time period. In addition, arguably this will force students to undertake some preparatory work for *all* of the possible questions, and so help to ensure a broad engagement with the module content.

Of course, not all plagiarism need be intentional, and we owe it to our students to ensure that they know how to reference correctly and are familiar with what plagiarism is and how to avoid it. This requires practice; for example, in creating references and identifying examples of plagiarised work. Encouraging students to use a computer-based checking system such as JISC's 'Turnitin' can also help them to recognise potential plagiarism and to develop their referencing skills. We need to be aware, too, that for some international students, there may be cultural barriers to understanding the concept of plagiarism; as Leathwood (2005) points out, *"what may be regarded as plagiarism in the UK could be seen as good practice in some other contexts"* (p316)*.*

e. Practical issues

This relates both to the 'efficiency' and 'effectiveness' of assessment. Efficiency requires that it should not be excessively onerous (in terms of time and effort) for either staff or students. It is common nowadays for course documentation to specify both the total hours of student learning which is deemed appropriate for a particular module /unit of study (depending on size), but also, the time to be spent on (summative) assessment. Tutors should try to ensure that the demands of their particular assessment do not exceed those specified. In addition, with ever increasing pressures on staff time and larger student cohorts, it is important that the tutor is capable of undertaking the marking within a reasonable time period (again, marking turnaround norms are often specified). This may be a particularly important issue if the module/unit is a School - or university-wide one, taken by hundreds of students, as students are, understandably, often anxious to receive prompt feedback; and prompt feedback can most effectively provide timely 'feed-forward'. There may also be issues around the practicality of the planned assessment activity itself, if it is likely to involve a large amount of staff time and/or what might be viewed as making excessive demands on other resources. Equally, of course, assessment must be effective – it must permit a fair and valid assessment of whether (and to what standard) the desired learning outcomes have been achieved.

Lines and Gammie (2004) – reported in Trotter (2006) – adopt an interesting approach to classify assessments according to both effectiveness and efficiency:

- Stars (which meet both criteria)
- Wots (a waste of time – they achieve neither criterion)
- Big Macs (efficient and fast but not effective), and
- Persian cats (which look good and are effective, but are too time consuming and hence inefficient)

It may well be illuminating to reflect on which category our favoured approaches to assessment belong!

Assessment criteria

The use of closely defined assessment criteria remains somewhat contentious amongst academics in higher education. Holroyd (2000) argues that this is for three related reasons:

i) that assessment decisions cannot be formulaic (or, as he calls it, use the 'tick-box' approach); "*the attachment of meaning and significance…is inescapably a judgement to be performed by persons"* (p37)

ii) that assessment judgements are uncertain and problematic, requiring "*inferences to be made…from more or less adequate evidence provided in samples of observable behaviour and artefacts / products"* (p38)

iii) that there may a temptation, in the context of increasing student numbers, to delegate assessment judgements to non-academic staff.

Ecclestone (1999) points out that there is a "*tension between learning outcomes to empower learners or to ensnare them in restricted forms of learning*" (p31), and "*It is already apparent in outcome-based systems that an instrumental attitude of just doing what is needed for assessment reduces the desire to do anything challenging or cognitively difficult"* (p47). So, by defining precise learning outcomes and associated assessment criteria, Ecclestone argues that this can encourage students inappropriately to narrow their learning aspirations, merely to ensure they can demonstrate the (potentially) narrow range of criteria specified. However, external pressures towards more explicit performance criteria in higher education (eg via HEQC (1997) and the Dearing Report (1997)) have resulted in the increasing pre-eminence of the criterion referenced approach to assessment. Hence, this guide recognises that just as there are now inescapable pressures on institutions to identify specific outcomes for all learning activities, there is generally a requirement for academic staff to use criterion-referenced approaches to the assessment of those learning outcomes. So, the issue which is focused on here is how best to design assessment criteria which can adequately assess students' achievement of the learning outcomes.

a. generic criteria

One common approach, to try to ensure consistency both between modules/units within and across courses, is to specify *generic* criteria, which relate to the general skills and competences which are expected to be demonstrated at each level of study; for example, the ability to analyse, synthesise, apply theory, structure an argument, etc. Some examples from Northumbria University are shown in Figures 1 (broad definitions by mark range), Figure 2 (specifying particular characteristics and skills, also by mark range), and Figure 3 (with a greater distinction between mark ranges).

Research by O'Donovan et al (2000) found that, although students found assessment grids of this type helpful in clarifying what was required by a piece of work, there were a number of criticisms which could be summarised as:

i) the vagueness and imprecision of the criteria, and

ii) the subjective interpretation of these criteria by staff

Together, these could lead to student disillusionment with the assessment criteria; as O'Donovan et al point out, *"negative experiences can dash expectations"* (p79) and so prove to be de-motivating for some students. They suggest a need for the criteria to be explained ('interpreted') for students in advance by tutors, which could also have the added advantage of helping to ensure more consistent interpretation and application by these tutors when using the criteria to assign grades / marks (see c. below).

Figure 1

Example of level 6 assessment criteria from the Newcastle Business School

	LEVEL SIX
First (80 – 100)	Exceptional scholarship for subject. Outstanding ability to apply, in the right measure, the skills necessary to achieve highly sophisticated and fluent challenges to received wisdom.
First (70 – 79)	Knowledge and understanding is comprehensive both as to breadth and depth. A mature ability to critically appreciate concepts and their inter-relationship is demonstrated. Clear evidence of independent thought. Presentation of work is fluent, focused and accurate.
Upper Second (60 – 69)	Knowledge base is up-to-date and relevant, but also may be broad or deep. Higher order critical appreciation skills are displayed. A significant ability to apply theory, concepts, ideas and their inter-relationship is illustrated
Lower Second (50 – 59)	Sound comprehension of topic. Reasoning and argument are generally relevant but not necessarily extensive. Awareness of concepts and critical appreciation are apparent, but the ability to conceptualise, and/or to apply theory is slightly limited.
Third (40 – 49)	Knowledge is adequate but limited and/or superficial. In the most part, description/assertion rather than argument or logical reasoning is used. Insufficient focus is evident in work presented.
(30 – 39)	Minimal awareness of subject area. Communication of knowledge frequently inarticulate and/or irrelevant.
(0 – 29)	Poor grasp of topic concepts or of awareness of what concepts are. Failure to apply relevant skills. Work is inarticulate and/or incomprehensible.

Figure 2

Level 6 grade expectations (Psychology)

	First (70% +)	Upper Second (60-69%)	Lower Second (50-59%)	Third (40-49%)	Fail (20-39%)	Bad Fail (0-19%)
Coverage of the question	Covers all aspects of the question.	Covers most aspects of the question.	May not address some major aspects of the question.	Fails to address a number of major aspects of the question.	Addresses relatively few of the major aspects of the question. May be too short.	Addresses none of the major aspects of the question. Probably too short.
Knowledge of relevant material	Evidence of extensive independent reading including books and recent journal articles (in addition to suggested readings).	Evidence of independent reading including books and journal articles.	Answer based mainly on lecture material.	Some relevant information from lectures.	Little evidence of relevant knowledge. May cite personal anecdote.	Almost no relevant knowledge. May rely on personal anecdote.
Accuracy	All the material is accurate.	There are no major factual errors.	There may be some minor factual errors.	There may be some major factual errors.	There may be many major factual errors.	Little or no factual accuracy
Relevance	All the material is directly relevant.	Almost all the material is directly relevant.	Some of the material may not be directly relevant.	Much of the material may not be directly relevant.	Little of the material is directly relevant.	Answers a totally different question to that set.

Clarity of expression	All points expressed clearly and succinctly.	Most points expressed clearly and succinctly.	Some points may not be expressed clearly.	Not always clear what was intended.	Often difficult to discern what was intended.	Hardly ever possible to discern what was intended
Organisation	Excellent (possibly original) organisation of the material.	Very clear organisation of material.	Clear organisation of material.	Some organisation of the material	Little structure apparent.	No structure apparent
Evaluation of theory, methodology and/or empirical evidence.	Shows excellent appreciation of the strengths and weaknesses of theories, methodologies and empirical evidence and their interplay. May show knowledge of the historical development of the field.	Shows good appreciation of the strengths and weaknesses of theories, methodologies and empirical evidence and their interplay. Perhaps some indication of the history of the area.	Makes some attempt to evaluate theories, methodologies and empirical evidence and to justify claims.	Assertion with little concern for evidence.	Assertion without concern for evidence.	Assertion without evidence

Personal Contribution	May present own (possibly novel) view of the material, perhaps integrating evidence from or drawing parallels with other areas of the discipline. May make insightful predictions about the future development of the area.	May present own view of the material, perhaps integrating evidence from or drawing parallels with other areas of the discipline. May make sensible predictions about the future development of the area.	May make some attempt to present own view of the material showing some concern for its justification.	May make some attempt to present own view of the material but with little concern for its justification.	May present own view of the material but without any attempt to justify it.	May present a personal view that is irrelevant to the question.

Source for Figures 1 and 2: Guidelines for Good Assessment Practice at Northumbria University, 2004

Figure 3

Taught postgraduate Programmes: generic assessment criteria

	Mark Range	Postgraduate Generic Assessment Criteria
Distinction	86-100	Exemplary work providing evidence of a complete or near complete grasp of the knowledge, understanding and skills appropriate to level 7. All learning outcomes met a high level. Exemplary in: use of primary sources of literature from a range of perspectives; development of analysis and structure of argument; critical evaluation of theories including those at 'cutting edge' of the discipline; creative original use of theory, research methods and findings; presentation of information to the intended audience.
	76-85	Outstanding work providing evidence to an extremely high level of the knowledge, understanding and skills appropriate to level 7. All learning outcomes met, most at high level. Outstanding in: use of primary sources of literature from a range of perspectives; development of analysis and structure of argument; critical evaluation of theories including those at 'cutting edge' of the discipline; creative use of theory, research methods and findings; presentation of information to the intended audience.
	70-75	Excellent work providing evidence to a very high level of the knowledge, understanding and skills appropriate to level 7. All learning outcomes met, many at high level. Excellent in: use of primary sources of literature from a range of perspectives; development of analysis and structure of argument; critical evaluation of theories including those at 'cutting edge' of the discipline; some creative use of theory, research methods and findings; presentation of information to the intended audience

Commendation	67-69	Very good work providing evidence of the knowledge, understanding and skills appropriate to level 7. All learning outcomes met, some at a high level. Very good in: use of up-to-date material from a variety of sources; development of analysis and structure of argument; critical evaluation of theory; application of relevant theory, research methods and findings to the problem in question; presentation of information to the intended audience
	63-66	Good work providing evidence of the knowledge, understanding and skills appropriate to level 7. All learning outcomes met, many are more than satisfied. Good in: use of up-to-date material from a variety of sources; development of analysis and structure of argument; critical evaluation of theory; application of relevant theory, research methods and findings to the problem in question; presentation of information to the intended audience
	60-62	Good work providing evidence of the knowledge, understanding and skills appropriate to level 7. All learning outcomes met, many are more than satisfied. Good in most of the following aspects: use of up-to-date material from a variety of sources; development of analysis and structure of argument; critical evaluation of theory; application of relevant theory, research methods and findings to the problem in question; presentation of information to the intended audience
Pass	57-59	Highly satisfactory work providing evidence of the knowledge, understanding and skills appropriate to level 7. All learning outcomes are met, some are more than satisfied. Highly satisfactory in: use of relevant material from a variety of sources; development of analysis and structure of argument; evaluation of theory; application of relevant theory, research methods and findings to the problem in question; presentation of information to the intended audience.
	53-56	Satisfactory work providing evidence of the knowledge, understanding and skills appropriate to level 7. All learning outcomes are met. Satisfactory in: use of relevant material from a variety of sources; development of analysis and structure of argument; evaluation of theory; application of relevant theory, research methods and findings to the problem in question; presentation of information to the intended audience.

	50-52	Acceptable work providing evidence of the knowledge, understanding and skills appropriate to level 7. All learning outcomes are met. Adequate in: use of relevant material from a variety of sources; development of analysis and structure of argument; evaluation of theory; application of relevant theory, research methods and findings to the problem in question; presentation of information to the intended audience.
Fail	45-49	Work is not acceptable in providing evidence of the knowledge, understanding and skills appropriate to level 7. A substantial majority of the learning outcomes are met, however, and the others are nearly satisfied. Adequate in most but not all of the following aspects: use of relevant material from a variety of sources; development of analysis and structure of argument; evaluation of theory; application of relevant theory, research methods and findings to the problem in question; presentation of information to the intended audience.
	30-44	Work is not acceptable in providing evidence of the knowledge, understanding and skills appropriate to level 7. Most of the learning outcomes are met, however, and many of the others are nearly satisfied. Adequate in at least some of the following aspects: use of relevant material from a variety of sources; development of analysis and structure of argument; evaluation of theory; application of relevant theory, research methods and findings to the problem in question; presentation of information to the intended audience
	1-29	Work is not acceptable and shows little evidence of the knowledge, understanding and skills appropriate to level 7. Few of the learning outcomes are met. Inadequate in several, or seriously inadequate in at least one of the following aspects: use of relevant material from a variety of sources; development of analysis and structure of argument; evaluation of theory; application of relevant theory, research methods and findings to the problem in question; presentation of information to the intended audience
	0	Work not submitted OR Work giving evidence of serious academic misconduct (subject to regulations in ARNA Appendix 1) OR Work showing no evidence of the knowledge, understanding and skills appropriate to level 7. None of the learning outcomes are met

Source: Dordoy A (2007), Academic Registry, Northumbria University

b. discipline and/or topic specific criteria

Alongside any generic criteria, course and module/unit tutors may wish to devise criteria specific to the topic(s) being assessed. It is not possible, within the scope of this guide, to explore particular examples of specific criteria, such is the possible breadth of different disciplines; however, in general, these will relate to the *content* of the assessment, and the specifics of the question asked, rather than to the academic and transferable skills demonstrated. For example, if the assessment concerns a public policy issue, criteria could identify whether the student has demonstrated an awareness of the underlying problems and issues, key policy development factors, implementation issues, and/or evaluation of the success of the policy against appropriate criteria (depending on the exact nature of the question). In contrast, for a design question, criteria might consider the accuracy and/or usefulness of the design drawings, and the interpretation of, and adherence to, the design brief. If we are to avoid the potential problems (explored earlier) of limiting students' approaches to learning, perhaps the key question when specifying specific assessment criteria must be, could the learning outcomes be demonstrated in ways which are not reflected in these criteria? And if so, how can the criteria be amended to encompass or permit these alternatives?

c. application of the criteria

The challenges for the tutor do not end once a set of assessment criteria has been devised for, as Holroyd (2000) argues, *"the notion that consistency problems in assessment are solved by the production of a set of assessment criteria is woefully simplistic"* (p35). He claims that assessors need to develop shared understanding of the meaning of the criteria in practice. Hand and Clewes' (2000) research at Nottingham Trent Business School (in relation to the assessment of dissertations) found "*a lack of consensus over what differentiates a 2.1 from a 2.2,* *particularly at the margin"* (p15), which lead them to suggest that "*we need to look much more closely at our construction of, and adherence to, marking guidelines if consistency is to be achieved and quality assured"* (p19). So, it is not sufficient simply to have developed a set of criteria; what is also critical is for all marking staff to *share* a clear understanding of the meaning and interpretation of those criteria, as this is fundamental to an equitable marking process – which is examined next.

Marking assessments

According to Holroyd (2000), assessment decisions require:

- Discipline-specific knowledge, which should be greater than that likely to be exhibited by those being marked
- Assessment craft knowledge, which is "*understanding of assessment that inevitably comes from experience of assessing and some, however minimal, reflection on it*" (p36)

However, he also argues the need for 'assessment scholarship', some understanding of the body of research and literature on assessment *"which at least can help assessors escape the obvious pitfalls and which at best can illuminate better practice"* (p36).

Hornby (2003) explored 'marking models' adopted by a range of staff at Aberdeen Business School, and found that all claimed to use a criterion-referenced model, "*where criteria were identified and in some cases given a weighting"* (p447). However, in practice, some groups were also likely to use a holistic model, marking 'intuitively' based on the work as a whole (for example, economists), whilst others were more likely to adopt a 'menu marking' model, awarding marks for each part of the assessment (for example, accountants). As Yorke et al (2000) point out, *"the more precisely assessment criteria can be specified,* *the easier it is to award marks for components of the overall performance"* (p20).

It is good practice to ensure moderation of marking standards, ideally through blind, double marking (meaning that a second assessor marks without any knowledge of the first assessor's judgement). However, given the sheer volume of work to be marked in universities today, and shortness of turnaround timescales in many cases, it is perhaps more common now to find work being sample-moderated. This usually involves a pre-defined sample being examined and the grading checked by a second assessor. This, of course, requires that the assessors share a common understanding of the application of the marking criteria, but *"it is clear that sharing standards is not easy"* (Price, 2005, p21). The existence of a 'model answer' or marking criteria or grade descriptors will not, of themselves, ensure similarity of interpretation. If there are only two assessors (such as the main marker and a second marker), it my be relatively easy for them to discuss and agree interpretations of criteria, but, where there are a number of markers, more formal approaches may have to be adopted. Holroyd (2000) points out that, as a result of modularisation, there has been a decline in the 'traditional' approach to assessment, in which *'the examiners know each other well..(and) there is the opportunity for shared understandings about assessment criteria and standards to develop*" (p30), though Price (2005) questions whether such close-knit marking communities have ever existed. Holroyd argues that *"the research evidence points to the importance of assessors being members of a communicating network"* (p31), which Price refers to as 'communities of practice' (p221). She identifies some of the methods by which understanding of standards can more effectively be shared, which include shared resources (such as grade descriptors, model answers, exemplars and sample marked scripts), 'marking bees' (at which everyone marks together and discusses samples), and discussion about marking and moderation to verify understandings. However, referring to research by Elander (2002), Price suggests that markers can only assimilate ‘a limited number of aspects of student work’, so limiting the number of aspects to be assessed may result in more reliable marking (p227).

It is common for module leaders to review average marks, the mark distribution and the range of marks in an effort to identify any anomalies in tutors’ marking. However, it cannot be assumed that any particular cohort will exhibit a ‘normal’ distribution of marks, so this needs to be reviewed in the light of cohort performance across all modules. In addition, Hornby (2003) claims that *“the type of marking system used can affect the distribution of marks”* (p444); specifically, grade scales will tend to produce a wider spread of marks than percentage scales. Hornby’s research identified very large unused ranges of marks in some modules, especially those which might be defined as ‘qualitative’, and found that, in almost two thirds of final year modules, “*over 50% of the percentage mark scale is not used”* (p440). This leads to ‘the problem of spurious precision’, in that, although the impression given is that marks represent ‘common currency with a common value’, in reality, they do not.

These differences in marks will, of course, influence final degree classifications and especially the percentage of ‘good’ degrees. In addition, as Simonite (2003) points out, *“systematic differences in the outcomes of assessment by different methods raise questions of fairness to students who aim for the same award, but follow programmes that differ in terms of the assessment methods used to measure performance”* (p460). Hence, tutors do need to be concerned about systematic and large differences in the mark ranges generated by different types of assessment. If markers can not be persuaded to adopt similar marking conventions (to generate similar mark distributions), course teams may need to consider whether re-scaling of the mark distributions for some modules (to fit the cohort's 'normal' distribution) is desirable.

Whenever possible, it is good practice to use anonymised marking of both exam scripts and courseworks, as assessors may be influenced, either positively or negatively (and not necessarily consciously), by their knowledge of the particular student. Where this is impossible, such as with project work that requires close staff supervision, then it may be appropriate to exclude the supervisor as an assessor. For events which cannot be repeated (for example, an oral presentation), then the use of a video recording might be considered, or, if not, at least there should always be two assessors to verify the judgement.

One final issue for consideration is whether the presentation of students' work can unfairly affect marking judgements. Of course, if presentation is part of the assessment criteria, then it is right that it should; however, there is a view that some types of presentation will subconsciously influence marking decisions, with features such as font size and type, line spacing, and margins suggested as relevant. Research by Hartley et al (2006) indicated that font size in particular seemed to have an impact on results, with essays using a 12 point font gaining, on average, 4.1% more marks than those using a 10 point font. Overall, the characteristics which apparently merited the highest marks were those with a Times Roman 12 point font, double spaced with unjustified text, and with a line space to denote new paragraphs, which achieved an average of 3.8% more than alternative combinations. This suggests that it may be fairer to students if course teams specify a standard presentation format, so that no student unwittingly selects a format which may, unintentionally, elicit a lower tutor evaluation.

Conclusion

This guide has highlighted that assessment plays a key role in helping students to learn and to develop their capacities as independent, lifelong learners, as well as in judging their achievements. There is growing emphasis on criterion-referenced (rather than norm-referenced) assessment, though this may be inappropriate for some forms of assessment. Planned assessments should be valid, reliable, consistent, diverse, efficient, understandable, support learning and provide effective feedback. Assessment methods need to be efficient and effective, appropriate for the type of knowledge or skills being tested, and aligned to the learning outcomes. When designing assessments, care should be taken to limit the opportunities for plagiarism and tutors should carefully monitor the marking outcomes for any adverse equality impacts. Generic criteria can help to promote marking consistency across modules and courses, but only if markers ensure that they share common interpretations and applications of the criteria. Finally, it must be recognised that large differences in the mark ranges arising from different types of assessment may impact on final degree classifications, so course teams may need to consider whether re-scaling of the distribution for some modules is desirable.

References

Broadfoot P (2000), preface in Filer A (ed), *Assessment: social practice and social product*, London, Routledge Falmer

Cranton P (2006), Rethinking Evaluation of Student Learning, *Higher Education Perspectives,* vol 2 no 1, Ontario Institute for Studies in Education, University of Toronto

Crooks T (1988), The impact of classroom evaluation practice on students, *Review of Educational Research,* vol 54 no 4, pp 38-481

Ecclestone K (1999), Empowering or Ensnaring? the implications of outcome based assessment in higher education, *Higher Education Quarterly,* vol 3 no 1, pp29-48

Elander J (2002), Developing aspect-specific assessment criteria for examination answers and coursework essays in psychology, Psychology Teaching Review, Vol 10 no 1, 31-51

Gibbs G (1999), Using Assessment Strategically to change the way students learn, in Brown S & Glasner A (eds), *Assessment Matters in Higher Education*, Buckingham, Society for Research into Higher Education / Open University Press

Gibbs G & Lucas L (1997), Coursework assessment, class size and student performance, *Journal of Further and Higher Education*, no 21, pp 183-192

Habermas J (1971), *Knowledge and Human Interests,* Boston, Beacon Press

Hand L & Clewes D (2000), Marking criteria: an investigation of the criteria used for assessing undergraduate dissertations in a business school, *Assessment and Evaluation in Higher Education*, vol 1 (1), pp7-27

Hartley J, Trueman M, Betts L & Brodie L (2006), What price presentation? The effects of typographic variables on essay grades, *Assessment and Evaluation in Higher Education*, vol 31, no5, pp523-534

HEQC (1997), *Assessment in Higher Education and the Role of 'Graduateness'*, London, HEQC

Holroyd C (2000), Are Assessors Professional?, *Active Learning in Higher Education*, London, Institute for Learning and Teaching in Higher Education and Sage Publications

Hornby W (2003), Assessing Using Grade-related Criteria: a single currency for universities?, *Assessment and Evaluation in Higher Education*, Vol 28, No4, pp435-454

Knight P (2006), The local practices of assessment, *Assessment and Evaluation in Higher Education, vol 31, No 4, pp 435-452*

Leathwood C (2005), Assessment Policy and Practice in Higher Education: purpose, standards and equity, *Assessment and Evaluation in Higher Education,* vol 30, No 3, pp 307-324

Lines D & Gammie E (2004), *Assessment Methods Report*, Education Committee of the International Federation of Accountants

McDowell L, Sambell K, Bazin V, Penlington R, Wakelin D, Wickes H and Smailes J, (2006), *Assessment for Leaning: current practice exemplars from the Centre for Excellence in Teaching and Learning*, Newcastle, MARCET Staff Development Resource Centre Red Guide

Maclennan E (2001), Assessment for Learning, *Assessment and Evaluation in Higher Education,* vol 26 no 4, pp 307-318

Madaus G F (1994), A technological and historical consideration of equity issues associated with the proposal to change the nation's testing policy, *Harvard Educational Review*, 64 (1), 31-48

Martin M (1997), Emotional and Cognitive effects of examination proximity in female and male students, *Oxford Review of Education,* Vol 23 no 4, pp 479-486

QAA (1996) *Understanding Academic Standards in Modular Frameworks*, London, QAA

O'Donovan B, Price M & Rust C (2000) The Student Experience of Criterion-Referenced Assessment, *Innovations in Education and Teaching International,* vo 39, no 1, pp 74-85

Pirie M (2001, January 20), How exams are fixed in favour of girls, *The Spectator,* pp 12-13

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

Simonite V (2003), The Impact of coursework on degree classifications and the performance of individual students, *Assessment and Evaluation in Higher Education,* vol 28 no 5, pp 449-470

Tan K H K & Prosser M (2004), Qualitatively different ways of differentiating student achievement: a phenomenograhic study of academics' conceptions of grade descriptors, *Assessment and Evaluation in Higher Education*, vol 29 no 3, pp 267-282

Trotter E (2006), Student perceptions of continuous summative assessment, *Assessment and Evaluation in Higher Education,* vol 31 no 5, pp505-521

University of Northumbria (2004), *Guidelines for Good Assessment Practice at the University of Northumbria,* Newcastle, University of Northumbria at Newcastle

Woodfield R, Earl-Novell S, Solomon L (2005), Gender and Mode of Assessment at university: should we assume female students are better suited to coursework and males to unseen examinations? *Assessment and Evaluation in Higher Education,* vol 30 no1, pp 35-50

Yorke M, Bridges P, and Woolf H (2000), Mark Distributions and Marking Practices in UK Higher Education, *Active Learning in Higher Education,* vol 1(1), 7-27