



Assessment Series No.7

A Briefing on Key Concepts Formative and summative, criterion & norm-referenced assessment

Peter Knight

Peter T Knight, once a historian, now works in the Department of Educational Research at Lancaster University. Although he continues to do research into and publish about schoolteaching (he co-authored *The Politics of Professionalism*, published by Continuum in 2000), he is more engaged with higher education matters. A long-standing interest in the assessment of learning, sparked by work on the use of multiple choice questions to assess adolescents' understanding of history, has been manifest in his writings on higher education. It should lead to the 2003 publication of *Assessment, Learning and Employability*, which will be written with Mantz Yorke.

Contents

Summary	3
Assessment purposes – summative and formative, feedout and feedback	3
Exhibit 1. Fifty assessment techniques	4
Exhibit 2. Extract from a social research methods course handbook	6
Table 1. Summative and formative assessment	9
Mixing summative and formative assessment purposes	11
Four assessment concepts: reliability, validity, affordability, usability	12
Reliability	12
Table 2. What are the features of reliable summative assessment?	13
Validity	14
Affordability	15
Usability	16
Norm-referencing and criterion-referencing	17
Norm-referencing	17
Table 3. Marks and grades in a norm-referenced system	17
Criterion-referencing	18
Exhibit 3. Potential advantages of criterion-referenced assessment	19
Exhibit 4. Qualifications to claims made for criteria-referencing	20
Four knotty issues	21
Plagiarism	21
The assessment of competence	22
Grade inflation	23
Aggregating marks	24
What teachers and programme leaders can do	25
Teachers	25
Programme leaders	26
Conclusion	27
References	28

Generic Centre Guides and Briefings

Welcome to the Learning and Teaching Support Network Generic Centre's series of Assessment Guides and Briefings. They aim to provide a series of overviews of important issues and practices in the field of assessment for the higher education community.

The Assessment Guides are intended for colleagues with particular roles and for students, as their titles suggest. The Briefings are primarily intended for lecturers and other staff involved in supporting learning.

The Assessment Series is a snapshot of a field in which development is likely to be rapid, and will be supplemented by specific case studies produced by the LTSN Subject Centres.

The series was developed by Brenda Smith and Richard Blackwell of the LTSN Generic Centre with the support of Professor Mantz Yorke. Experts in the field were commissioned for each title to ensure that the series would be authoritative. Authors were invited to approach the issue in their own way and no attempt was made to impose a uniform template.

The series editors are grateful to colleagues in LTSN Subject Centres and other senior colleagues who refereed the series, and of course to the authors for enabling its publication.

We hope that you will enjoy the Assessment Series and find it interesting and thought-provoking. We welcome your feedback and any suggestions you may have for future work in the area of assessment.

Professor Brenda Smith
Head, LTSN Generic Centre

Richard Blackwell, Senior Adviser,
LTSN Generic Centre

Professor Mantz Yorke,
Liverpool John Moores University

November 2001

Summary

This paper explains a set of key assessment terms in order to help teachers draw up good assessment plans for their courses. Not only is it a useful guide to often-misunderstood concepts, it also advises on things to avoid, things to do with caution and things to do with enthusiasm. Finally, it contains suggestions for heads of department who increasingly have to think of programme-wide assessment systems. They need to understand the language in order to design well.

Assessment purposes – summative and formative, feedout and feedback

As Exhibit 1 indicates, there are many interesting assessment methods and there is no shortage of books on how to use them (for example, Brown and Knight, 1994; Banta et al., 1996; Hounsell *et al.*, 1996; Brown et al., 1997; Walvoord and Anderson, 1998; Heywood, 2000). However, in order to decide which methods are best for your course you need to have a fairly well-formed sense of what you are trying to achieve in setting the assessment tasks. That is not to imply that some assessment methods are fit for some assessment purposes and not for others. Assessment methods are neutral. However, it is easier to see how the different methods might best be used if you have a clear idea of what each assessment task is for.

The Quality Assurance Agency expects that things valued enough to be stated as course learning outcomes will be assessed, so the first piece of clarification should concentrate on those learning outcomes. Exhibit 2 contains the learning outcomes for a social research methods course and shows how they are picked up by assessment tasks. What you see in the left-hand column of the table is the result of the teacher's deliberation about which assessment methods are best fitted to the course learning outcomes. Good course design means drawing up a matrix like this to show the purposes that each assessment task serves.

Notice also that five of the assessment methods have the suffix **(S)** and one is followed by **(F)**. **S** stands for summative and **F** for formative and they describe a second set of assessment purposes.

Assessments are frequently used to sum up a person's achievement. In these cases, there is a *summative* purpose behind the tasks learners are set. Summative assessment provides 'feedout', in the shape of a warrant to achievement or competence (such as a degree certificate), and in the form of information that can be used as performance indicators in appraising the work of teachers, departments, colleges and national systems of education. Assessment for summative purposes is high-stakes assessment. One implication is that those being assessed are likely to do all they can to conceal ignorance and suggest competence. That is in some tension with another implication, namely that when the purposes are summative then the assessment should get it right - should be accurate, objective and reliable. Of course what should be, what can be, and what can be afforded are not necessarily the same, a theme to be developed later in this paper. These issues are elaborated in the next section. For the present it is enough to notice that assessment for summative purposes is one approach to assessment and that it carries with it the need to take reliability very seriously.

Exhibit 1. Fifty assessment techniques
[Based on Box 9.1 in Knight, 2002b]

1. Artefacts/ products, especially in fashion, design, engineering, etc.
2. Assessment as gatekeeping: students gain entry to classes only on production of bullet point summaries etc.
3. Assessment banks. Students have access to a question & answer bank. They learn how to answer all of them but are assessed on a sample).
4. Assessment of work-based learning (in a variety of ways, many times, by a variety of people, for different purposes).
5. Book, website or program reviews.
6. Classroom assessment techniques. They are *brief* tasks that tell the teacher something about the *class's* grasp of the material. (See Angelo and Cross, 1993).
7. Completing structured summaries of readings, debates etc.
8. Computer-based self-assessment.
9. Concept maps. Excellent way of seeing how students understand complex content and relationships.
10. Contribution to threaded electronic discussions.
11. Defence of lab records.
12. Design and build (similar to 2, above).
13. Dissertations and theses.
14. Electronic monitoring of web searches, program use & communications.
15. Essay writing - one 5000 word, piece (make harder/easier by varying amount of tutorial guidance, range of reading expected, novelty of the topic/problem, time available, conceptual complexity, etc.).
16. Essays writing - 2 x 2500 word pieces.
17. Exhibitions of work, posters, products. History students have curated museum exhibitions in lieu of doing a dissertation.
18. Field work and lab work assessment (traditional and well established).
19. Formative assessment of logs/journals/portfolios (when the purposes are formative, students identify areas for discussion. *If* summative, sampling within the logs etc. is recommended, especially if students know in advance the areas that are likely to get closest attention).
20. Games and simulations.
21. 'General' assessments, drawing together learning in several modules.
22. Making annotated bibliographies for next year's students.
23. Making models (literally, in some subjects, conceptual models in others).
24. Multiple choice questions (they do not have to be only tests of information, although it is a lot quicker to write MCQs like that. See also 3, above).
25. New tests in which learners use old software/programmes/notes.

26. Objective Structured Clinical Examination (OSCE).
27. Open-book, end of course exams.
28. Orals and vivas.
29. Peer assessment. Some try to use it summatively but it's a lot easier if done for formative purposes.
30. Performances. Vital in the assessment of competence. Note massive problems assessing complex performances fairly and reliably. Simulations sometimes possible.
31. Personal response assessments. Usually done in classes where each student has an electronic response pad. Teachers ask questions and they press a key to show their answer. Can be used for classroom assessment or test purposes.
32. Posters.
33. Production of structured logs of project/dissertation progress and reflection on it.
34. Projects.
35. 'Real' problem working, which involves defining 'fuzzy' situations, bringing some order to ill-defined issues, analysing the problem and suggesting solutions.
36. Replication of published inquiries.
37. Role-playing.
38. Self-assessment. Skill at self-evaluation is valued by many employers, which is a reason for having self-assessments. Easiest when used formatively.
39. Seminar presentations (in or out of role; with or without use of video, OHT, PowerPoint, etc.).
40. Short answer questions. (MCQs *plus* some explanation of the thinking: limit to 100 words per response?)
41. [Short] appraisals of target papers.
42. Small-scale research or enquiry.
43. Statements of relevance, which are short pieces of writing, 1000 words, perhaps, making claims about the relevance of a workshop, article, field observation etc. to another task or activity. See Bourner, O'Hare and Barlow (2000).
44. Submission of claims to achievement with reference to portfolio (if this is to be summative, I suggest grading on the claim alone *provided that* sufficient evidence supports the claims).
45. Takeaway papers/questions/tests.
46. Terminal, unseen examinations and other individual time-constrained assignments.
47. Two part assessments. Elements of a task are formatively assessed but the final product is summatively assessed.
48. Web page creation.
49. Writing exams/tests/assessments to tutor specification.
50. Writing memoranda or journalistic summaries.

Exhibit 2. Extract from a social research methods course handbook

Outcomes

By the end of the course, you should:

1. Demonstrate knowledge of mainstream educational and social research methods-
see Programme Specification 10.2B (2).
2. Be able to critically engage with issues concerning the relationships between research and knowledge, and with the fitness of different research methods for different purposes-
see Programme Specification 10.1 (2).
3. Have skill at reading and evaluating research reports –
see Programme Specification 10.2A (1), 10.2B (1).
4. Be able to design a feasible, small-scale research inquiry-
see Programme Specification 10.2B (1, 2).10.2C (3).
5. Take responsibility for organising and managing much of your own learning –
see Programme Specification 10.2C (2, 4).
6. Work effectively with others, both to their benefit and yours-
see Programme Specification 10.2C (5, 6).
7. Treat the Internet as a mainstream learning resource –
see Programme Specification 10.2B (3).
8. Present your conclusions orally to an audience-
see Programme Specification 10.2B (6).

This table links these outcomes to the main pieces of work you will do.

Assessment tasks	1	2	3	4	5	6	7	8
Coursework 1 – write a 200 word literature review (S)	✓	✓	✓		✓	✓	✓	
Coursework 2 – designing a research inquiry (S)	✓	✓		✓	✓	✓	✓	✓
Coursework 3 – write an evaluation of a published research paper (S)	✓	✓	✓		✓	✓	✓	
Submit bullet point lists related to each week's set readings. (F)	✓				✓		✓	
Examination – Q.1 note making (S)	✓							
Examination – Q2 designing an inquiry (S)	✓	✓		✓				

Assessments are also used to identify what learners need to do in order to improve their work. This second approach to assessment, which is intended to inform students about how to do better, is often called *formative* assessment. Any task that creates feedback to students about their learning achievements can be called formative assessment. Diagnostic assessment, which involves using carefully designed tasks to try and identify barriers to learning, can be seen as a type of formative assessment that is little used in higher education. Notice that formative assessment, with its emphasis on providing useful feedback, is more helpful when learners are open about their limitations and don't try to conceal ignorance or bury mistakes. Whereas summative assessment purposes discourage learners from being open, formative assessment purposes thrive on disclosure. Furthermore, with formative assessment the stakes are not so high - no-one's future rides on the accuracy of advice about continuing to improve your work - which means that we need not worry so much about reliability. Reliability is obviously good if you can get it or can afford to get it but it is not central to formative assessment in the way that it must be central to summative assessment. That is very useful because there are some things which we want students to learn that cannot be reliably assessed or cannot be reliably assessed *with the resources available*.

Look again at learning outcomes No.5 and No.6 in Exhibit 2. I'm not sure whether it is possible to assess No.5 reliably (except perhaps through simulations or in a laboratory and then I'd be quite unsure exactly what was being assessed) and there are practical and ethical problems with No.6 as well. However, if the purpose is formative, to create feedback to students about No.5 and No.6, then it is easy to see plenty of authentic ways of doing it, especially if everyone understands that the feedback comes from thoughtful appreciation's of the student's work *and* that the formative assessment will not produce a reliable measure of student achievement. The feedback will be fuzzy in comparison with the products of summative assessment, which may be reported to two decimal places, but it will be reaching learning achievements that elude summative assessment, a claim I shall develop in the next section of this paper.

It is very useful to be able to assess for formative reasons because curricula in higher education are giving increasing prominence to complex learning outcomes and to 'soft skills' - they are claiming to foster inter-personal skill, emotional intelligence, creativity, critical thinking, reflectiveness, incremental self-theories, autonomy and such like. Later in this paper I will explain why it is hard *and* expensive, if not simply impossible, to assess *reliably* a complex achievement such as critical thinking or a 'soft skill' such as emotional intelligence. Here I assert that to be the case and conclude that it is very convenient that we have an alternative in the shape of formative assessment which generates useful feedback that affords learners the chance to inform their future learning.

It might seem that assessing with formative purposes in mind is being puffed-up only because 'real' assessments, with their tests, certainty and two decimal places, are currently not up to the job of measuring some voguish learning outcomes. Leave to one side my claim that this is not a problem of technique because reliable assessments are just not possible, or not affordably possible. Formative assessment is much more than something to do when outcomes have to be assessed and there's no reliable way of producing summative data. A review of 681 research publications on formative

assessment showed '... conclusively that formative assessment does improve learning' and argued that if best practices were achieved in mathematics on a nationwide scale that would raise 'average' countries such as England and the USA into the top five (1998: 61); the possible effect size of 0.7 is '... amongst the largest ever recorded for educational interventions' (Black and Wiliam, 1998: 61). There is a good case for seeing formative assessment as an extremely powerful contributor to student learning. You could almost say that makes it *more* important than summative assessment. Naturally, this promise is not to be attained without meeting a number of conditions. For example, Black (1998) concluded that formative assessment is best when it relates to clear criteria and where the comments are not accompanied by marks or grades. Brown and Knight (1994) identified more than a dozen beliefs and practices that were needed if formative assessment were to work in higher education and Torrance and Pryor (1999) have confirmed the point that good formative assessment can make a considerable difference to the quality of learning.

But in the words of Black and Wiliam, 'It is hard to see how any innovation in formative assessment can be treated as a marginal change in classroom work' (1998: 16). Good formative assessment means designing learning sequences that afford plenty of opportunities for good learning conversations arising from feedback on good tasks that are matched to course learning outcomes. Good formative assessment therefore implies thinking about learning, teaching and assessment, not just about assessment. At the same time it opens up many possibilities. Consider, for example, peer- and self-assessment. The educational case for each is strong (Boud, 1985). However, attempts to use them have been dogged by fears that they are unreliable and will breed academic malpractice. Many complex solutions have been suggested to these problems (Lejk and Wyvill, 2001; Li, 2001). Things become much more straightforward if peer and self assessment are used formatively. Here we also see the beginning of an answer to the charge that formative assessment is resource-intensive. It is if feedback is always created by teachers. However, when shrewd assessment planning exploits the potential of formative peer- and self-assessment, and when course teams have taken other steps to make it easier to give rapid feedback (Knight, 2002, Chapter 9), then lowstakes assessment can save time.

Four of the conclusions that can be drawn from this review are:

- Assessment does not mean testing, although tests and examinations are one way of assessing.
- Assessment does not have to be summative.
- Some things, especially complex achievements and 'soft skills' might best be formatively assessed.
- Assessment plans are necessary and, ideally, should be in course handbooks given to students and on the course website that they use.

Table 1 provides a more detailed summary of the differences between the two assessment purposes.

Table 1. Summative and formative assessment

Dimensions of difference	Assessment as measurement (Summative)	Assessment as judgement (Formative)
Ontology (theory of what exists) and epistemology (theory of how we know about it)	Common sense view that there is a reality that is readily known through diligent use of 'scientific' methods.	There is a problematic relationship between what may exist and what is known. There is not, therefore, any right way to the truth.
Assumptions about achievements	Achievements are seen as transferable. Good measurements predict achievements in other times and contexts.	There is only a limited transfer of learning so there can be no strong claim about learner's performance in other contexts. <i>Assessment data are not good predictors.</i>
Typical products	'Feedout' in the shape of warrants to achievement	Feedback in the shape of improvement 'conversations'.
Priorities	<ol style="list-style-type: none"> 1. Reliable measures of achievement 2. Motivating learners 3. Providing information to guide learning 	<ol style="list-style-type: none"> 1. Providing comments that afford opportunities for better learning 2. Motivating learners
Treatment of complex human learning achievements	Reductionist. Assumes that complex achievements can be separated into component parts that can be reliably assessed. The total score is then treated as a valid measure of complex achievement.	Complexity has emergent properties, which means that the whole is more than the sum of its parts. Complex achievements must be judged as they are.
What achievements are most likely to be assessed in this way?	Understandings and performances that can be fairly captured by low-inference judgement methods – convergent, routine, lower-order achievements.	Complex achievements – divergent creations, non-routine judgements, 'soft skills'.

Dimensions of difference	Assessment as measurement (Summative)	Assessment as judgement (Formative)
Common assessment techniques	Fixed-response tests of all sorts, tests of routine application of formulae, methods etc. Objective methods Objective observers	High-inference judgements of authentic achievements on projects, in work placements etc. Assessment <i>is</i> inference Fair methods Fair assessors
How are assessments communicated	Often in numerical form	Often in words - narratively. Numbers sometimes used as shorthand.
Suggestions for improvement	Use programme-wide assessment plans to identify what is going to be summatively assessed and when. Invest in doing those summative assessments well. Develop assessment criteria for use in summative assessments. Ensure that there are repeated observations of learning outcomes that are summatively assessed. Use multiple observers/ assessors. Be considerate in reporting achievement – as far as possible, use plain English to say what has been achieved.	Throughout the programme use learning indicators as points of reference in assessment conversations. Ensure that there are pervasive programme-wide messages about the significance of formative assessment. Formative assessment takes time – design it into the assessment plan. Teach students how to do formative assessment, peer- and self-assessment in particular. Link formative assessment with student claims to employability.

Mixing summative and formative assessment purposes

An obvious question to ask at this point is whether formative and summative purposes are mutually exclusive: can an assessment task be set for formative and summative purposes? It is tempting to think that the best effects would come from setting a formative task that also produced reliable marks for summative purposes. However, most experts in the field think it is best not to try this. The reason is that when the stakes are low, people are prepared to disclose things that they try to hide when the stakes are high. With formative assessment, we really want to know what people have difficulty with so that we can help them - we want them to disclose what they can't do. If they think that an assignment will also have a summative purpose, they won't disclose their problems, which means that the formative intention will be ruined. Elton (1996) has suggested that until students are comfortable with summative assessment they may find it hard to learn with understanding, rather than memorisation, in mind, a point which Mantz Yorke endorses in his Assessment Series paper *A Guide for Senior Managers*.

However, you can have formative and summative assessment in the same course. For example, use multiple choice questions (MCQs) for summative assessments of information and set a series of 100 word summary questions as formative assessments of understanding. Two-part assessment involves setting an assignment, then looking at what students produce and making two or three suggestions for improving performance on a similar task in future (formative). Summative, high-stakes assessment could follow: for example, set all students one question based on the assignment and their reports, which is to be answered in 45 minutes (say). Or set 'stepped tasks' where there are compulsory reporting points where students get feedback on some elements of a larger task which is then summatively assessed.

High-stakes assessments can give students feedback that has real formative value. Arguably, all summative assessments should provide it and some departments have marksheet templates that instruct tutors to make one or two suggestions about how the student could best improve the next piece of work. They should be written with programme learning indicators or criteria in mind, although sometimes improvement-focused feedback needs to encourage students to reflect on the way they go about their work (in other words, to stimulate metacognition) which requires comments that will not usually be based on programme learning criteria. Students often say that they do not get improvement-centred feedback; that feedback from summative work can take weeks or months to come; and that they do not get feedback from examinations that is intended to help them become better at doing examinations.

Four assessment concepts: reliability, validity, affordability, usability

The distinction between summative and formative purposes calls for an account of four key concepts in assessment. You might recognise them as key concepts in social research, which is not surprising since social research and assessment both explore human thought and action and both have to consider the ways in which we might know about them (Knight, 2001, Chapter 2). Table 1 may be about assessment but it would not take much to make it into a comparison of two approaches to social research.

I have gone easy on the references in my coverage of these four concepts, partly because the four intermingle in ways that makes it unhelpful to give references to each separately. If you want to pursue these issues, overviews are provided by Brown et al., (1997), Broadfoot (2000), Heywood (2000), Linn (2000), Black (2001) and Knight (2001, 2002a).

Reliability

We depend upon reliable measures of real, observer-independent phenomena, which we expect to be:

1. Objective, in the sense that observer biases have not compromised them.
2. Accurate, in the sense that the methods of measurement are stable and sensitive. A rubber band is neither a stable instrument for measuring height and a weighbridge is not a sensitive measure of a dieter's weekly weight loss (or gain).
3. Repeatable. The measurement procedures must be clear and consistent from case to case. If they are not, then the same measurement process is not being used and we might expect a consequent change in the results, simply as a result of variation in procedures.
4. Analytically sound. We expect tests to be correctly analysed (marked) and the results to be accurately entered into an appropriate software package that runs the right statistical routines.

Although these routines might produce 'objective' data, they often fail, say their critics, to reflect the complexity of human achievements. 'Objective' data, they argue, tend to be artificial and to give information about those things that can be objectively measured. Anything else, no matter how significant, gets ignored. Consider the case of attempts to measure school effectiveness. The data used to judge school effectiveness are exam results. They are quite reliable but their apparent objectivity comes from examining students only on those aspects of school subjects that can be reliably graded. In every subject students' skill at designing, managing and completing semi-independent inquiries goes unmeasured by examinations, even though it is a part of the curriculum. The exams are equally insensitive to other widely-valued qualities, skills and achievements (Black 2001).

Table 2 elaborates on these issues. Sometimes, though, these simplifying tendencies are acceptable, particularly where it can be shown that a reliable test is a good predictor of complex behaviours. That claim is often made for IQ tests and other standardized tests

Table 2. What are the features of reliable summative assessment?

Measurement theory	Implication for the reliable assessment of learning
That which is to be measured must exist.	Obvious. But do skills really exist or are they just convenient names to refer to a variety of human practices and achievements? Even if skills exist, can we say they are stable? In other words, how confident can we be that a good performance on this task in this context today means that we will perform well on a different task in a different setting in three weeks time?
We need to be able to produce a valid definition of what exists to allow us to identify what we should be measuring and what we ought not to measure.	If we can't define it, we can't measure it. The reason why there is such argument about what 'critical thinking' means is that unless a definition can be agreed we can't measure it because we can't agree what it is. It may not be possible to define that which we want to measure. For example, intelligence has been defined as what intelligence tests measure. Just because something is important, it doesn't mean that we can agree what it is.
It is usual to measure samples of that which is to be measured. We need to take enough, representative samples to be sure that we get a fair picture.	Two points here. The more samples that we measure, the greater the reliability. If we want to make statements about someone's critical thinking power, we would want a lot of measurements taken in different contexts with different tasks. In practice we often over-generalise from a few measurements done in similar settings on similar tasks.
The instruments and scales must be reliable.	In practice, it's unusual to know that the instruments are reliable.
Instruments should also be valid.	Validity is often a bigger problem. Take reading, for example. We often use word recognition tests to measure someone's reading level. But if reading is defined as a complex thinking process (and it often is), these tests have low validity. Whatever they're measuring, it isn't reading.
The observer needs to be reliable.	Assessors need to be trained and monitored. Second and third markers should be used.
Data must be reliably analysed.	They must all use grading criteria and the grading criteria must be useable.
Numerical data from different sources should not be merged unless we believe that they all describe the same thing.	This is a technical and little-known point. By all means combine numerical data from different sources but don't imagine that you've got anything more than a set of numbers. It is hard, perhaps impossible, to say what the numbers you get <i>mean</i> .
The final measurement should have predictive validity, which means that the scores should correlate well with those obtained by other means.	The most expensive and thorough way of measuring complex human achievement is the assessment centre. Even so, Hunt (1991), a world expert on organisational leadership, reckoned that only 10% or 20% of variance in promotion could be accounted for by assessment centre scores. In other words, assessment practices often correlate poorly with the real-life achievements that we would expect them to predict.

widely used in the USA. Even here, though, there is a counter view that simple tests best measure skill at doing simple tests and that performance on them does not correlate well with later professional and career success (Sternberg, 1997). The debate has little relevance for most higher education teachers because the reliability of these North American tests is the outcome of considerable development work far beyond the reach of most academic departments. In fact it is sobering to compare the characteristics of tolerably-reliable 'A' level examinations and most attempts at reliable assessment in higher education – higher education practices look ramshackle in comparison.

Validity

Let's say that validity is about assessing the things that the programme specification and course learning outcomes say that you value. Who would assess anything different? Most of us do - unwillingly and perhaps unwittingly too. I will concentrate on just one reason, although there are others. The reason is that the more we can simplify what we are trying to assess the more reliable the assessment. Simplicity breeds reliability and reliability craves simplicity (Breland, 1999; Linn, 2000). When students are learning well-defined material which is full of unambiguous information it is easy to see how their retention of it could be very reliably assessed. But what if we need to assess a complex, fuzzy, ill-defined achievement, such as critical thinking? I recognise that 'critical thinking' has several meanings and feel a lot easier if we talk about it with reference to a subject area, a topic and a purpose. But even as we try to define it we are excluding other meanings and beginning a process of simplification that continues as we then begin thinking about objectivity, accuracy, repeatability and sound analytical procedures. We might end up with a test of critical thinking that can be reliably used with any university student but which, as a result, has to be content free so that it looks very much like a cross between an IQ test and a logic test. So, we often do not assess validly because the simplification that happens when we try to assess, *especially* when we try to assess summatively and for high-stakes purposes, compromises the validity of the assessment.

Is that valid? When we say that someone is good at critical thinking do we really mean that they score highly when solving puzzles under test conditions? The need for reliability pushes us towards certainty and simplicity but modern higher education curricula value complex, fuzzy achievements exemplified by soft skills, autonomy, creativity, incremental self-theories, interpersonal fluency, etc.

What matters most, reliability or validity? For a long time the answer was reliability, the thinking being that no result could be valid if it was not reliable. As far as determinate learning outcomes, such as information recall, are concerned, it is quite reasonable to plump for reliability because the pursuit of reliability does not significantly harm the thing that you are trying to assess - it is a straightforward, determinate sort of achievement, quite suited to reliable assessment. But where complex and ill-defined learning outcomes are concerned, putting reliability first shatters validity in two ways. First, the tests simplify complexity and provide information about something quite different than that which you think you are assessing. Secondly, because reliability is associated with summative assessment purposes, which are high-stakes assessments, students tend to engage with the learning outcomes as *simplified for assessment purposes*.

That does not have to happen when assessment has formative purposes. Feedback is likely to relate to fuzzy performance indicators but it, and the conversations that accompany it, do not need to be regulated and ritualised in ways that hinder people from talking about the complex learning outcomes that are supposed to be course priorities.

Affordability

If reliability tends to simplify, the tendency can be resisted and it is possible, as those concerned with vocational qualifications have shown, to get acceptable levels of reliability when assessing complex, authentic performances. Suppose we have to get tolerably-reliable judgements of clinical competence or skill at making oral presentations. Assume that we can agree on definitions of clinical competence or presentational skill and on what would count as good evidence. We then need to have multiple observations of performance by different trained observers. With complex skills it is likely that we would want these performances to be fair samples of the range of possible performances, which might have the effect of increasing the overall number to be observed. Breland (1999) showed that it takes three trained judges looking at three pieces of writing to make reliable assessments of students' skill at writing one page pieces on general, unprepared topics. More complex achievements would need more evidence and more judges to attain the same levels of reliability. Fleming (1999) showed how many sources of bias, or unreliability, there are in the grading of something as familiar as academic essays. He, like Breland, was suggesting that reliability could be improved and that it is costly and complicated; costly – both financially and in terms of opportunity costs – to train assessors; to develop good grade indicators or other judgement guides; to provide second, even third markers; and to make multiple observations. Yet it is obviously complicated to schedule multiple assessment opportunities and it might even be undesirable in short courses where it could seem to be little teaching, just assessment.

The conclusion that tolerable levels of reliability can be achieved is of little help to most teachers because they cannot afford to try and do it in their courses, especially if theirs are bitesized, modularised, semesterised or unitized courses. Teachers interested in assessment issues often ask for advice on improving reliability and are crestfallen to find that they cannot afford to act on it: they lack time, skill and other resources. However, this does not force us into the preposterous position of suggesting that higher education can only produce tolerably reliable judgements of low-level, achievements such as information recall. The answer to this and many other assessment problems is a programme-level answer, dependent on leadership and systemic thinking. It may be impossible to get reliable judgements of skill at oral presentation out of one module but it is not hard to see how they could be had from a programme-wide approach to assessment. (Programme specifications should identify the learning outcomes and the learning, teaching and assessment practices that are orchestrated to go with them.)

What the individual teacher cannot afford, let alone manage, programmes can. Good leadership and systemic thinking are the keys to many assessment problems that conscientious teachers assume they have to solve alone. They can't.

Usability

Although I cannot say a lot in this briefing paper about improving the usability of assessment procedures (it is a complex topic), I am mentioning it for the sake of completeness. (It is also not a topic you will find much about in the standard works on measurement theory.) There are two other reasons why it does not get sustained attention here. First, because *internal usability*, the usefulness of assessment to students (in the shape of formative assessment feedback), teachers and academic managers, is related to the quality of thinking about assessment purposes, reliability, validity and affordability. To put it another way, unless there are plenty of opportunities for formative assessment, there is little point in thinking about the usability of assessment judgements.

Secondly, *external usability*, which is particularly about feedout and the usefulness of assessment information to postgraduate schools and employers, is a programme and/or institutional matter, not something where individual teachers have the leading role. The Quality Assurance Agency expects higher education to improve the external usability of assessment information by contributing academic transcripts within student progress files. Unfortunately, the samples they have provided (<http://www.qaa.ac.uk/crntwork/progfileHE/guidelines/progfile2001.pdf>) do not seem fit for the purpose (see Adelman, 1990, on good practice and academic transcripts).

Good assessments should be useable. Individuals cannot fix systemic shortcomings in the usability of assessments.

Norm-referencing and criterion-referencing

Whether the purposes of an assessment task are summative or formative, the question of standards has to be considered? What points of reference are to be used?

Norm-referencing

Norm-referencing is one answer. The idea is that assessments do not compare student achievement directly to statements of learning outcomes but give data that allow us to rank achievements, comparing one student to another. Norm-referencing is comparative, telling us that *this* student is better than another, similar to a third and not as good as a fourth. It is normative in the sense that the grades awarded reflect students' rank order and expectations about the proportion of A's, B's, C's, etc. that it is reasonable to award. Table 3 shows the grades awarded to two cohorts of students. It is clear that (i) it is rank, not mark which determines the grade and (ii) that the distribution of grades is kept fairly constant, reflecting underlying beliefs about how achievements ought to be truly distributed. Notice that norm-referencing is more a way of treating marks than a system for awarding them.

The National Curriculum in English schools involves assessing children against attainment targets, which are statements of eight levels of understanding that children might display in their work on the subjects they study. Although this system looks to be criterion-referenced, because grades, in the form of judgements about levels of achievement, are to be related to performance criteria, the official expectation that 50% of 11 year olds will be at level 4, about 20% at level 3, 20% at level 5, and so on, introduces norm-referencing into the system. In many departments it is traditional to resolve uncertainties about assessment marks by norm-referencing, when the big issue becomes one of where to draw the grade boundaries, for as Table 3 showed, a mark that in one year gets an 'A' grade gets a 'B' in another. This process is seen as the last defence against rampant grade inflation, which is covered in more detail in the next section. This desire to maintain standards may be praiseworthy but there are severe objections to norm-referencing as a way of doing it. The three most important are that it is insensitive to changes in student learning, that it is uninformative and that it is based on a bad grasp of measurement theory.

Table 3. Marks and grades in a norm-referenced system

1999-2000		2000-2001			
Raw score range	Ranks	Grades	Raw score range	Ranks	Grades
68-81	1-9	A	67-93	1-10	A
61-67	10-35	B	59-66	11-34	B
50-60	36-69	C	50-66	35-65	C
44-49	70-85	D	41-49	66-82	D
36-43	85-93	E	37-40	83-90	E
<36	93-100	F	<37	91-99	F

Many of us believe that good teaching makes a difference to the quality of student learning and hope that it shows up in the form of better marks. We object strongly if we are told that we must not give 'too many' good marks because that would 'devalue our standards'. This use of norm-referencing appears to be designed to deny that good students, good teaching and good learning can make much difference, if any, to the outcomes of an assessment process whose results are broadly known in advance. This is wrong. Worse, all anyone can deduce from the results is that someone with an 'A' grade was in the top ten percent of a cohort in that particular higher education institution. It says nothing about what they have achieved, apart from being in the top decile of a cohort of unknown quality, and gives no basis for comparison with students elsewhere. The process is also theoretically deficient. It may be true that we can establish how stable attributes are distributed amongst a population and on that basis make fairly reliable predictions about their distribution in good-sized samples drawn at random from that population. If 10% of the population have a characteristic then we can be quite confident that the same proportion will hold in a sizeable sample drawn at random from the population. However, norm-referencing only works under certain conditions and, as the following list shows, these conditions are not usually met when it comes to student learning in higher education:

1. The attributes in question – understanding and 'skills' - are not as stable as is sometimes assumed. Indeed the whole point of a programme is to improve student understanding or performance by means of superb teaching and learning arrangements.
2. We neither know how these attributes are distributed in the population of students taking the subject nor, given that they are not stable attributes, could we find out.
3. A cohort in one university is not a random sample of all the students taking that subject in UK universities. In fact admissions procedures are designed to ensure that it is not. It would be a mistake, then, to assume that we know what the distribution of achievement within it ought to be.
4. Even if the above points are ignored (and they shouldn't be), a cohort is often too small for it to be safe to make inferences about the distribution of the attribute within it.

Criterion-referencing

Criterion-referencing is distinct from norm-referencing in that it is a system for awarding marks and it is fundamentally opposed to norm-referencing in that it is open to situations in which everyone fails (rotten teaching, lazy learners, bad assessment) or everyone gets grade 'A'. It is widely used, especially in assessments of competence. For example, the QAA subject benchmarks (<http://www.qaa.ac.uk/crntwork/benchmark/benchmarking.htm>) are written as criteria identifying achievements commensurate with different levels of performance. The theory is simple. Identify what counts as successful performance or good attainment, specify it precisely and judge evidence of achievement accordingly. In most assessment situations levels of achievement are described, each with their own criterion or level descriptor. When achievements are complex it is likely that there will be several criteria or descriptors for each level and when complex performances are being assessed, as in the assessment of classroom teaching or performance in a law moot, then assessors will be simultaneously judging multiple achievements against multiple criteria.

Criterion-referencing is widely preferred to norm-referencing. Exhibit 3 lists seven possible advantages, the most important of which is that it involves identifying what learners need to understand or be able to do without making any assumption about the percentage who will succeed. However, the seven advantages are over-stated. Agreed, criterion-referencing tends towards each of the seven but for reasons deeply seated in philosophy and psychology (Knight, 2001, Table 2.1) it cannot completely deliver on any of these seven counts. Exhibit 4 elaborates the reasons for being more cautious about criterion-referencing than government agencies and practitioners often are. Criterion-referenced assessments are inherently complex and as higher education tries to do more than impart information, it becomes harder still to conceptualise learning goals and it takes longer to capture them in useful criteria statements. Once fixed on paper, these statements then have to become shared, understood and applied in common ways, which is also highly time consuming. Because they usually involve high-inference judgements, they take longer to use than low-inference marking schemes and second and even third assessors are called for. And the data they produce are often less wieldy than a simple but misleading numerical score. Reliability can be enhanced but at a cost in terms of time, money and lost opportunities to do other things.

It is worth remembering here that a lot of the concerns about the incipient unreliability and expense of criterion-referenced assessment become less serious if the assessment purposes are formative. In this case the criteria, even quite ambiguous, loose or fuzzy criteria, which I prefer to call 'indicators', provide a good basis for assessment conversations intended to identify fruitful ways of working to improve achievements.

Exhibit 3. Potential advantages of criterion-referenced assessment

1. Assessment criteria clearly identify what is valued in a curriculum.
2. In criterion-referenced curricula teachers know exactly what they should teach.
3. Level descriptors make it clear to learners what they have to show in order to get a particular mark.
4. Level descriptors make it possible to give learners feedback which identifies what they need to do in order to get better marks.
5. Level descriptors can be used to make assessment feedout informative, identifying exactly what learners have achieved.
6. Criteria and descriptors make it possible to give feedback on complex work with some reliability. They provide agreed standards expressed in a shared language that together serve as points of reference for markers to use when trying to put marks to students' work. Although there is a (misguided) view that good criteria carefully used by well-trained assessors will eliminate reliability problems, there is no doubt that they reduce the degree of unreliability in assessments of complex achievements.
7. It is possible to make judgements about the quality and quantity of learning.

The fine distinctions, certainty and final language that go with summative assessment purposes are not needed. Indeed, in formative assessment ambiguity can be more useful than certainty – ambiguity invites thought and discussion whereas clear-cut judgements by those in authority tend to discourage discussion and encourage outward compliance. (See Moshman, 1999 for a good discussion of learning in situations where one person is powerful and the others are not).

Exhibit 4. Qualifications to claims made for criteria-referencing

1. It is very difficult to develop statements of learning outcomes, criteria for awards at different levels, and formulations of thresholds. For example, there are problems of conceptualisation. What exactly is involved in, for example, 'critical thinking'? Disputes about this and other complex learning goals flourish and agreement is unlikely. What I call leadership you call paternalism; what you call good oral communication, I call bombast.
2. Attempts to produce precise criteria lead to a proliferation of sentences and clauses culminating in complex, hard-to-manage documents. However, the obvious alternative, settling for fewer, looser statements, combines manageability with imprecision, thereby undercutting the rationale for criteria-referencing, namely that students should know exactly what they have to do, as should academic staff and other stakeholders.
3. It is important to remember that these are criterion-referenced judgements, not criteria-determined ones. The criteria are reference points in processes of judgement, aids not replacements. In my department we talk of grade indicators, not of grade criteria. Any claim that criteria can or should displace judgement is met with a rapid rebuttal citing Krippel, Wittgenstein and psychologists too many to name.
4. Even the most carefully drafted criteria statements have to be translated into concrete and situation-specific terms. For example, what exactly is 'clear communication', or 'speaks with enthusiasm', or 'mastery of one common word-processing package' The research evidence makes it quite clear that there is a degree to which criteria cannot be unambiguously specified but are subject to social processes by which meanings are contested and constructed (Greatorex, 1998; Price and Rust, 1998).
5. It takes a lot of training before assessors feel comfortable with criterion-referenced marking and do it in broadly similar ways. As with all high-inference assessments, reliability is expensive.
6. Criterion-referenced assessments do not necessarily produce useable data. They are no worse than norm-referenced assessments when data from different criteria levels are converted to numbers and aggregated to produce a single score. However, when the aim is to retain the richness of the assessment data the result can be bloated transcripts and references, which are often full of jargon and only understandable to insiders.
7. Because criteria referencing makes it possible to assess the full range of programme aims it invites attempts to assess personal qualities. That, in turn, raises ethical issues, including the fear that assessment may become a normalizing force in the colonisation of the affective domain.

Four knotty issues

In this section I comment on four of the assessment problems that worry a lot of teachers. They are by no means the only ones. They are presented partly because the suggestions for tackling them may be helpful in their own right but also because my suggestions show that many assessment problems are really leadership and system problems (Knight and Trowler, 2001). That is not to say that teachers should not care about them, nor to deny that there are things they can do. However, it is to say, unequivocally, that some of the things that individuals try to resolve are soluble by teams in the context of whole-programme planning and of institution-wide policy-making.

Plagiarism

Poor assessment practices invite plagiarism, although even good assessment practices can be compromised by sharp practices. Four ways to deter plagiarism are:

1. Prefer formative to summative assessments. When assessment has formative ends, then plagiarists only harm themselves. Remember, good formative assessment runs on disclosure, not on deceit, such as plagiarism.
2. Make more use of time-constrained individual assessments. Exams are the most famous example of this sort of assessment but there are lots of creative ways of setting individual tasks to be done under supervision. There are limits to what can be validly assessed in this way but there are also plenty of creative possibilities.
3. Don't set the same tasks year in year out.
4. Set distinctive tasks – for example comparing two recent papers, writing a review of a recent publication, making an annotated bibliography. The chances of plagiarism from outside the class are small and plagiarism within the class can be fairly easily identified.
5. Let students know that you are prepared to run suspect text through software that tells you if it is unlikely that a piece of writing is in the same style as the student usually uses. (There is a review of plagiarism detection sites on <http://www.coastal.edu/library/mills3.htm>)

In many subjects plagiarism says more about the quality of thinking than it does about students' moral failings. Plagiarism often shows students responding intelligently to teachers' slack assessment practices.

The assessment of competence

Teachers try to make reliable assessments of competence and, even if they recognize sources of unreliability, they generally feel fairly relaxed when traditional methods are used to assess traditional competencies – when essays are used to assess skill at constructing historical arguments. Concern grows when using assessment methods that are new to them, such as observation, to judge nontraditional achievements, like skill at working in groups. I suggest three approaches:

1. Don't try to get reliable assessments of achievements that tend to resist affordable, tolerably-reliable assessments. Do create plenty of opportunities for learners to get feedback on work that is related to fuzzy grade indicators (formative assessment) and encourage them to develop their own claims to achievement and to amass substantiating evidence.
2. If you have to summatively assess complex performances, then use the best grade indicators or observation schedule you can; try to get a colleague to help you assess and/or use peer- and self-assessments as supplementary evidence; and assess performance on several occasions, the more the better.
3. Ideally the measures described above would be written into a programme-wide assessment plan, so that by the time students came to graduate there would be many measures of performance by different observers all using the language of a common observation schedule or set of assessment criteria.

This third, systemic, programme-level approach is necessary when reliability is important. It is also good when combined with the first approach, since in that way students get a great deal of improvement-oriented feedback. The second approach is the least desirable because, in truth, it is near impossible to get much reliability in the assessment of complex achievements in a one-semester course.

The major point I want to make in these comments is that the assessment of competence is no great problem with formative and low-stakes approaches to assessment. When done for summative, high-stakes purposes, it is a problem that individual teachers should not be expected to tackle by themselves. When the purposes are summative, the problems are primarily systemic leadership problems.

Heywood (2000) is good on these issues, advocating a multi-strategy approach to the assessment of complex achievements.

Grade inflation

Students throughout education systems are getting steadily higher marks (Elton, 1998), although the tide is stronger in some subjects than others. There are broadly three explanations of what is often described as 'grade inflation', which is hardly an unbiased way of describing the phenomenon.

1. Some say that students are smarter nowadays (and recent evidence that IQ scores have risen by a fifth, on average, since World War II might substantiate that explanation).
2. Others say that the tests have got easier and assessors have effectively abandoned attempts to maintain the golden standards of yesteryear.
3. A third group argues that today's assessments use different methods and address a wider range of achievements than did the assessments of twenty-five years ago. It so happens that these modern assessments are more sensitive to things that English students do well and less sensitive to those that cause them difficulty.

This is an issue for debate and although research can contribute, well-known technical difficulties with using assessment data to make statements about changes in standards over time (Desforges, 1989), it will not resolve the argument. This means that it is not clear whether there really is a problem (there is only cause for celebration if the first explanation is correct), let alone what might count as a solution.

Furthermore, if there is a problem, it is not something that individuals, or even programme teams, can do much about. In the UK the QAA has tried to clarify standards with its benchmarking exercise (<http://www.qaa.ac.uk/crntwork/benchmark/benchmarking.htm>), although remarks in the last section suggest that there are limits to what can be achieved by specifying levels of achievement in the form of level descriptors.

At programme and course levels, the best safeguards against grade 'creep' are known, understood and used criteria. They will not prevent creep, especially where teachers keep getting better, but they are the only alternatives to norm-referencing, which may simultaneously prevent creep and any recognition of learning.

Aggregating marks

Assessments are not all made on the same scale even if, at first sight, it appears that the same scale is being used. It is well known that some departments use the mark range more freely than others. As a historian I hardly dared use a mark greater than 70 but in other subjects 70s are commonplace. Even within a subject that spans the Arts and Sciences, such as Economics or Psychology, there may be some courses using a range that effectively runs from 15% to 85% while others group the marks in the range 30%-73%. How can marks based on such different grading practices be fairly combined? For example, students in some subjects have far better chances of getting higher mean marks than those in others, simply as an artefact of the marking systems in use. Where degree classification procedures are sensitive to mean marks, some students have an advantage because of the subject they are studying. This can be particularly awkward in combined degrees where students get marks from Arts and Natural Science modules. Some universities have tried to get around this problem by insisting that all departments use letter grades which are benchmarked to degree classification levels, so that grade 'A' is a first class mark, B is an upper second. A short marking scale can result, which is why other universities insist that departments use a 20 point scale in which 20 is an outstanding first, 17 a borderline one and so on and so forth. While there is no doubt that this reduces the scale of the problem of different subjects using different scales, it does not quite solve it. This can be illustrated by noticing that something similar can also be seen *within* courses that use a variety of assessment methods in judging a range of achievements. I teach an undergraduate course which has three assignments: an oral group presentation, graded through indicators that direct attention to *presentational* qualities; an individual bibliographic search task, graded with reference to indicators relating to search, evaluation and selection strategies; and an individual report, graded more or less as a regular piece of analytical and critical argumentation. Students regularly get the best marks for the first assignment and the worst for the third. Is it good practice just to calculate the mean of the three marks, even though it involves aggregating unlike with unlike and produces marks for the module that are slightly but distinctly better than they would have been had the students done the traditional three essay/report questions?

When it is put like this it is fairly clear that there are enduring difficulties with an assumption that underpins most assessment practices in the UK, the assumption that information about different achievements, which are appropriately measured in different ways, can be put together and summarised by a single mark or grade which signifies academic achievement. That assumption and the operations that follow from it are not valid. They may be convenient but they are not valid. They may provide neat degree classifications but those classifications have no known meaning: they are floating signifiers, nothing more.

There are answers to problems like these: relying more on formative assessment and being less ambitious in what we try to catch with summative assessment; or reporting achievements that have been summatively assessed separately and as statements of achievement, not as mean scores. Individual tutors can choose to tinker with the first suggestion but really both answers need to be treated as policy problems, not as problems for individuals to work at in an *ad hoc* way.

What teachers and programme leaders can do

A theme of this paper has been that caring teachers get concerned about assessment problems that they cannot ease. They are system- institution- and programme-level matters, calling for leadership and systemic action. An appreciation of that helps teachers concentrate their energies where they can make a difference.

Within this Assessment Series there is a Guide for Senior Managers by Professor Yorke which extends the suggestions I make about what programme leaders can do, if they have the support of heads of department who are prepared to take seriously the role conflicts that make it difficult for academic staff to give as much attention as they might wish to teaching, learning, assessment and curriculum. Knight and Trowler (2001, especially Chapter 5) have a lot to say about approaches to these and other problems that vex department leaders.

Teachers

1. Reconsider the balance between formative and summative assessment purposes.
2. Consider extending the range of assessment methods.
3. Appropriate some learning criteria (subject benchmark statements will do) as a way of beginning to extend your thinking about what is being rewarded when you assess students and how marks relate to different levels of performance. Expect to revise them and to continue to revise them.
4. Network. LTSN subject centres, professional associations and other interest groups in this country and overseas are good sources of ideas to borrow and customise.
5. Amidst this activity hold on to the idea that many of the assessment problems you would like to solve are either (i) not solvable or (ii) most sensibly tackled at system level. Teachers are prone to feel guilt (Hargreaves, 1994). Sometimes it is helpful to say that guilt is not appropriate because solutions lie outwith the teachers' power.

Programme leaders

1. Make programme assessment practices a priority for departmental attention over a three year period, initially.
2. Review the amount of assessment on an assessment programme, looking at the range of methods and the balance of formative and summative. It is usual to find considerable imbalances (too many essays on history programmes, for example).
3. Look for consultancy help on the design and management of assessment systems. There is a place for educational development in the form of workshops on topics of interest but there is, I suggest, a massive and unmet need for educational development as consultancy. If programme leaders and heads of department do not press for consultancy, then the current skew towards workshop services for individuals will not be corrected.
4. Get some programme-wide criteria in place to help thinking about assessment. Concentrate on identifying the sorts of performance associated with, say, a lower second class degree. People usually exaggerate what they expect of students at any level, so you might check that the performance indicators that you write are not really descriptors of 2:1 achievements. Don't take them too seriously but treat them like a 'starter culture', a way of developing conversations about what is involved in assessing learning. They can be the beginnings of a common language to use about assessment.

Conclusion

This paper has outlined some of the meanings of fundamental assessment concepts and pointed to some implications for practice. Sometimes this has led to suggestions that individual teachers can apply to what they do but often the claim has been that practice falls irredeemably short of that which is necessary in order to come up to the standards expected by measurement theory (and the psychology and philosophy behind it). The suggestion is that in order to come up to those standards teachers, working in programme teams, need to be laying programme-wide assessment plans that encourage good learning. This improves their chances of passing fairly trustworthy judgements on those learning achievements that the team chooses to invest in assessing summatively and reliably. That view is outside the scope of this briefing but I have developed it elsewhere (Knight 2000).

References

Adelman, C. (Ed.) (1990) *A college course map: taxonomy and transcript data*. Washington: US Government Printing Office.

Angelo, T. A. and Cross, K. P. (1993) *Classroom assessment techniques: a handbook for college teachers*. San Francisco: Jossey Bass.

Banta, T. et al. (Eds.) (1996) *Assessment in practice*. San Francisco: Jossey Bass.

Black, P. (1998) Learning, league tables and national assessment. *Oxford Review of Education*, **24**(1), 57-68.

Black, P. (2001) Dreams, strategies and systems: portraits of assessment past, present and future. *Assessment in Education*, **8**(1), 65-85.

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

Boud, D. (1995) *Enhancing learning through selfassessment*. London: Kogan Page.

Bourner, T., O'Hare, S. and Barlow, J. (2000) Only connect; facilitating reflective learning with statements of relevance. *Innovations in Education and Training International*. **37**(1), 68-75.

Breland, H. M. (1999) From 2 to 3Rs: the expanding use of writing in admissions. In: Messick, S. J. (ed.) (1999) *Assessment in Higher Education: Issues of access, quality, student development and public policy*. Mahwah NJ: Lawrence Erlbaum Associates. pp.91-111.

Broadfoot, P. (2000) Assessment and intuition. In: Atkinson, T. and Claxton, G. (eds.) (2000) *The Intuitive Practitioner: On the value of not always knowing what one is doing*. Buckingham: Open University Press. pp.199-219.

Brown, G., Bull, J. and Pendlebury, M. (1997) *Assessing Student Learning in Higher Education*. London: Routledge.

Brown, S. and Knight, P. (1994) *Assessing Learners in Higher Education*. London: Kogan Page.

Desforges, C. (1989) *Testing and Assessment*. London: Cassell.

Elton, L. (1996) Strategies to enhance student motivation: a conceptual analysis. *Studies in Higher Education*. **21**(1), 57-68.

Elton, L. (1998) Are degree standards going up, down or sideways?. *Studies in Higher Education*. **23**(1) 35-42.

Fleming, N. (1999) Biases in marking students' written work. In: Brown, S. and Glasner, A. (eds.) *Assessment Matters in Higher Education*. Buckingham: Society for Research into Higher Education & Open University Press. pp. 83-92, .

- Greatorex, J. (1999) Generic descriptors: a health check. *Quality in Higher Education*. **5**(2) 155-166.
- Hargreaves, A. (1994) *Changing Teachers, Changing Times*. London: Cassell.
- Heywood, J. (2000) *Assessment in Higher Education*. London: Jessica Kingsley Publishers.
- Hounsell, D., McCulloch, M. and Scott, M. (eds.) (1996) *The ASSHE Inventory*. Edinburgh: University of Edinburgh and Napier University.
- Hunt, J. G. (1991) *Leadership: a new synthesis*. Newbury Park, CA: Sage.
- Knight, P. T. (2002a, forthcoming) Summative assessment in higher education: practices in disarray. *Studies in Higher Education*. **27**(2).
- Knight, P. T. (2002b, forthcoming) *Being a Teacher in Higher Education*. Buckingham: Society for Research in Higher Education & Open University Press.
- Knight, P. T. (2001) *Small-scale research*. London: Sage Publications.
- 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. and Trowler, P. R. (2001) *Departmental Leadership in Higher Education*. Buckingham: Society for Research in Higher Education & Open University Press.
- Lejk, M. and Wyvill, M. (2001) Peer assessment of contributions to a group project. *Assessment and Evaluation in Higher Education*. **26**(1), 61-72.
- Li, L. K. Y. (2001) Some refinements on peer assessments of group project. *Assessment and Evaluation in Higher Education*. **26**(1), 5-18.
- Linn, R. (2000) Assessments and accountability. *Educational Researcher*. **29**(2), 4-16.
- Moshman, D. (1999) *Adolescent Psychological Development*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Price, M. and Rust, C. (1999) The experience of introducing a common criteria assessment grid across an academic department. *Quality in Higher Education*. **5**(2), 133-144.
- Sternberg, R. J. (1997) *Successful Intelligence*. New York: Plume.
- Torrance, H. and Pryor, J. (1998) *Investigative formative assessment: teaching, learning and assessment in the classroom*. Buckingham: Open University Press.
- Walvoord, B. E. and Anderson, V. J. (1998) *Effective Grading: A tool for learning and assessment*. San Francisco: Jossey-Bass.

The Learning and Teaching Support Network Generic Centre

The Learning and Teaching Support Network (LTSN) is a network of 24 Subject Centres, based in higher education institutions throughout the UK, and a Generic Centre, based in York, offering generic information and expertise on learning and teaching issues that cross subject boundaries. It aims to promote high quality learning and teaching through the development and transfer of good practice in all subject disciplines, and to provide a 'one-stop shop' of learning and teaching resources for the HE community.

The Generic Centre, in partnership with other organisations, will broker information and knowledge to facilitate a more co-ordinated approach to enhancing learning and teaching. It will:

- Work with the Subject Centres to maximize the potential of the network;
- Work in partnership to identify and respond to key priorities within the HE community;
- Facilitate access to the development of information, expertise and resources to develop new understandings about learning and teaching.

The LTSN Generic Centre Assessment Series Guides for:

Senior Managers
Heads of Department
Lecturers
Students

Briefings:

Assessment issues arising from key skills
Assessment of portfolios
Key concepts: formative and summative, criterion and norm-referenced assessment
Assessing disabled students
Self, peer and group assessment
Plagiarism
Work-based learning
Assessment of large groups

Published by

Learning and Teaching Support Network (LTSN)

For more information, contact the Generic Centre at:

The Network Centre, Innovation Close, York Science Park, Heslington, York, YO10 5ZF

Tel: 01904 754555 Fax: 01904 754599

Email: gcnquiries@ltsn.ac.uk

www.ltsn.ac.uk/genericcentre