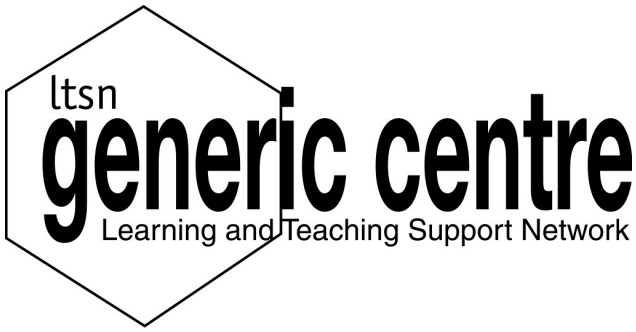


Assessment in Universities: a critical review of research

Lewis Elton

University College London

and

Brenda Johnston

CHERI, The Open University

Acknowledgements

We would like to express our thanks to the Generic Centre of the LTSN network which funded this review project. We would especially like to express thanks to our contact person in the Generic Centre, Professor Brenda Smith, who provided us with support throughout the process,

Lewis Elton and Brenda Johnston

Today we have the marking of folders

Today we have the marking of folders. Yesterday
We had assessments. And tomorrow morning
We shall have what to do after GCSE. But today
Today we have the marking of folders. Daffodils
Dance in their jocund glee around my garden,
And today we have the marking of folders.

This is the replacement mark-scheme. And this
Is the official mark-sheet, whose use you will see,
When you have read the grade descriptions. And this is the green book,
Which in your case you have not got. The tulips
Hold in the garden their silent eloquent gestures,
Which in our case we often also make.

These are the objectives, which are always to be observed
In setting all the assignments. And please do not let me
See anyone fiddling his marks. You can do it quite easy
If you have any brains in your head. The flowers
Are fragile and motionless, never letting anyone see
Any of them fiddling the marks.

And these you can see are the grade divisions. The purpose of these
Is to sort out the candidates. We can slide these
Rapidly backwards and forwards; we call this
Marking the units. And rapidly backwards and forwards
The early bees are assaulting and fumbling the flowers;
They call it marking the work.

They call it marking the work: It is perfectly easy
If you have any brains in your head; like the scheme
And the sheets and the syllabus and the new green book.
Which in our case we have not got; and the cherry blossom
Silent in all the gardens and the bees going backwards and forwards,
For today we have marking of folders.

Anne Anderton
(With apologies to Henry Reed)

Reprinted from the Times Educational Supplement, 13. 5. 1988.

Chapter 1
Setting the Scene
Lewis Elton and Brenda Johnston

Introduction

The Learning and Teaching Support Network (LTSN) is concerned both with improving current practice and with innovating, ie with doing things better as well as doing better things. One of the purposes of this paper is to give assessment an importance in this approach comparable to that of teaching. It does not have this at present, although it has long been known (e.g. Snyder 1970 and Becker 1968) that the nature of the assessment in a course has a profound effect on way that students learn. Hence a crucial aspect of a successful teaching and learning system is student assessment. It is no exaggeration to say that

> ***'If you want to change student learning then change the methods of assessment'***
> (Brown et al 1997, p 9).

The corollary of this is that if one changes the method of teaching, but keeps the assessment unchanged, one is very likely to fail. Thus to get the assessment right is vitally important.

Assessment must be in line with the desired learning objectives. This would appear to be an obvious criterion, if it were not a fact that in practice it is frequently not the case. How to get the assessment right, i.e. in line with the learning objectives or at least as right as possible, should be a major concern in any approach to assessment.

In view of the very large amount of research in the field of assessment, we have set ourselves the task of selectively, but critically, reviewing existing research, the selection being governed by our perception of the relevance and importance of the research. It would also appear that traditional assessment, whichever form this may take in different disciplines, is often accepted rather uncritically, which is much less the case for innovative assessment.

We will attempt to deal with three potential audiences to this paper:

1. Those who wish to maintain good practice within current assessment methods. We intend the 'Guide to Good Practice' for them (see Appendix to chapter 2).
2. Those who wish in the main to improve practice within current assessment methods. We intend chapter 2 for them.
3. Those who wish in the main to change current assessment methods. We intend chapter 3 for them.

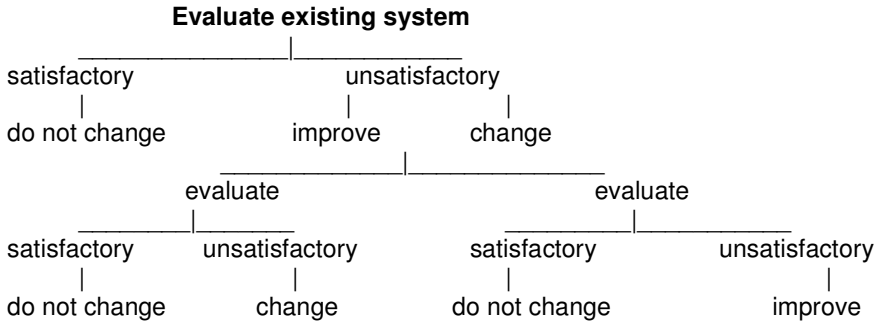
We hope however that members of all three groups will study the whole paper in order to appreciate the range of views presented. They may also find that what is traditional in one discipline may be considered innovative in others. We also hope that they may appreciate that most of what is called 'innovative', was already known to some students in the heady days of 1968. Thus two students from those days, A Powell and B Butterworth (1972) concluded their well argued critique of the then existing situation, with their preferred alternative:

> "The university system most worthy of experimental adoption, as we think, is that involving the abolition of all grades, and even of the pass/fail line. Syllabuses thus could atrophy, since their main purpose is to standardise learning in order to make assessment more rational. Lecturers would then be free to teach anything that students wanted to hear and talk about, and students would be able to transcend the traditional categories of learning. Certificates could inform anybody interested that so-and-so had attended whatever courses and had produced a certain volume of work, and beyond that each person would have to be judged purely on their merits for whatever job they applied for. It would of course be open to students, when applying for any job, to present samples of relevant work they had done at university, just as, at present, artists use portfolios of their work."

The issue of employer love of grades for easy sifting will be addressed in chapter 2.[1]

Doing things better and doing better things

In the past decades, educationists have been much more concerned with new methods of assessment (doing better things) than with improving current methods (doing things better). Both are required, but when one thinks of the vast number of timed examinations that go on in universities, then a small improvement in timed examinations might in the present situation have a far bigger effect than the replacement of the timed examination by a superior method of assessing, if this is practised by only a small minority. On the other hand, if an assessment method should be found to be intrinsically ineffective, then it should be changed before an attempt is then made to improve the new method. What all this means is expressed in the following diagram:



Whether to improve before trying to change is a matter of judgment which has to take into account most or all of the features of the system.

We will be concerned with assessment methods related to traditional as well as innovative teaching. With traditional teaching, there are of course accepted traditional assessment methods and we will investigate how far these may need improving and whether this can be done within existing regulatory frameworks. If they are in principle unsatisfactory, they may need changing, in which case they will then very likely conflict with the existing regulatory framework. While many universities (not all!) have relaxed their regulations in order to fit in with new forms of assessment, the universal requirement that in the end the results of all the assessments for a student have to be conflated into a single degree class continues to create a straight jacket which may have to be challenged. If the teaching is innovative, it will therefore often be difficult or even impossible to fit the appropriate assessment methods within the existing regulatory framework. If it is only difficult, then it is wise to be flexible, since changing entrenched regulatory frameworks is rarely easy. (Incidentally, entrenched regulatory frameworks can prove surprisingly flexible if handled with subtlety rather than force.) If it proves impossible to change the existing regulatory framework, it is usually wiser to abandon the attempt at a teaching innovation than to innovate and then be sabotaged by a totally inappropriate assessment system. It is therefore vital to link all assessment to the objectives which it is designed to assess and then to go through the steps indicated in the diagram above.

Good practice

The improvement of assessment practices requires both fundamental research - grounded in appropriate learning theories, and evaluative research - which evaluates both traditional and innovative practices. This might, although rarely does, take the form of action research, where results are fed back into the system in order to improve it, and which is then re-evaluated, thereby creating a virtuous spiral.

It must be stressed that much well established practice is far from good practice, in the light of such research. However an attempt at establishing a body of good practice for a very traditional institution, albeit within the constraints of the traditional conceptions of good practice, was made recently by one of us (see the Appendix to chapter 2) at University College London. It is worth noting that one of its constraints was that it had to fit existing Registry regulations!

[1] See *The predictive aspect of assessment* p.12, and *Degree class* p.18-19

The organisation of this Report

Following this jointly written Chapter One, the next two chapters will be written separately by LE and BJ respectively, with the final chapter again a joint effort.

Chapter Two, 'Challenges to established practice', raises questions about many "traditional" assessment practices. It documents the persistence of a largely unreflective traditionalism in spite of the existence of proven innovations and much relevant research over the past 35 years. One result of that research, which will be discussed, is that for university examinations - in contrast to psychometric testing - it is possible to choose the extent to which reliability or validity should be dominant in a particular assessment. This outcome, which is relevant to all forms of assessment, is indeed beginning to become generally accepted within universities, although pressures from outside increasingly seem to demand high reliability at the cost of validity. I then go on to a list of assumptions from a 1967 paper which at the time were unreflectively accepted and which have remained essentially unreflectively accepted ever since by the majority of practising examiners. The assumptions have however been researched into, and the outcomes of this research will be discussed. In the process, the concepts of positivism and interpretivism are introduced, which are essential to a deeper understanding of all assessment. Their basic characteristics are explained below and they are treated more fully in the following chapter.

Chapter Three, 'Some basic assessment dilemmas with particular reference to portfolios', focuses on links between underlying philosophical and theoretical positions, and empirical research on assessment. It performs three basic tasks.

1) It questions the traditional philosophical and theoretical bases of the summative assessment debate which has centred around concepts of "reliability" and "validity". This debate has largely been based within the positivist paradigm, with its belief in "objectivity", "scientific measurement" and certainty. Within this debate, assessment is viewed as a form of scientific enquiry to discover a student's ability, which is perceived as "a fixed, consistent and acontextual human trait" (Huot 1996, p.550). Much of academia has moved on from such certainties in the light of postmodernist and other challenges. Chapter Three attempts to apply such challenges to assessment and presents an interpretivist alternative which has been suggested in the literature. This approach perceives assessment as a socially constructed and socially contextualised practice. The purpose of this first part of the chapter is not to present either position as the answer to all assessment problems, but rather to raise awareness of the principles and practical implications underlying assessment practices.
2) The chapter then explores the implications of these different approaches to assessment through the concrete example of portfolio assessment. The chapter reviews research on agreement over outcomes in portfolio summative assessment in this light. It also points out gaps in the research.
3) In the third section, the chapter changes tack and moves to formative assessment and learning theories, again through the vehicle of portfolio assessment. It explores claims made that portfolio assessment can help students learn more effectively.

Chapter Four presents our overall conclusions, largely in the form of points to which we wish draw attention.

Chapter 2
Challenges to established practice

Lewis Elton

"It is now thirty years since serious doubts were raised about examinations, yet despite the fact that there has been no serious shortage of critics since then, very little has changed."
R Cox (1967a) p. 352.

"For many years I taught in universities. . . . I marked thousands of scripts without examining what the scripts could teach me about my capacity as a teacher and examiner."
E Ashby (1984), p. V.

"Something like 90% of a typical university degree depends on unseen time-constrained written examinations, and tutor-marked essays and/or reports."
P Race (2001), p 5.

The Persistence of Traditionalism

One aim of this chapter is to argue that assessment is still pervaded by a largely unreflective traditionalism, exemplified for instance in the quotations above, and that, as Race (2001) has indicated, truly innovative assessment is still very much a minority practice. A second aim therefore is to influence the majority, ie those who confine themselves to more traditional practices, to reflect on and perhaps change their present practices.

The first of the quotations comes from the June 1967 issue of Universities Quarterly which was devoted to 'Examining in Universities'. All the articles in it are still worth reading. In a particularly far sighted article, Oppenheim et al (1967) list 21 assumptions commonly made at the time in university examinations and then say: "Empirical testing is possible for most of these assumptions, but it will take a long time if it is done at all. In the meantime, the rationality of the examination system could be surely enhanced if universities were to decide which of these assumptions, here made explicit, they wish to maintain and which to abandon." To the best of my knowledge, few if any of these assumptions have been investigated, although, as apparently Race (2001) found out, the practices persist. His figure of 90% presumably is a rational guess, since relevant quantitative research does not exist. Atkins et al (1993), p.26 - 27, list six serious and persistent flaws in current assessment practice:

- No consistency in criteria used between subjects, within subjects, between institutions and within institutions for awarding of degree class
- Certain frames of reference which lecturers bring to assessment are systematically biased, but the bias is often subconscious and unrecognised
- Internally, lecturers have little idea of how others set and mark assignments; external examiners are not usually part of the curriculum design team, both are usually untrained in assessment
- Few lecturers understand the technical design factors which can affect assessment outcomes
- New forms of assessment , eg continuous assessment, are as prone to distortion as formal examinations
- Although there are exceptions, in many departments the approach to assessment remains conservative through ignorance.

Toohey (1999), pp 48 – 69, discusses the very different philosophical bases available for the design of courses and their learning objectives, and it may be expected that this information should in some way be contained in University prospectuses. University prospectuses usually do describe their course offerings and their scrutiny might therefore reveal the proportion of courses described purely in terms of disciplinary content without any reference to learning objectives, let alone philosophical bases. An interesting reflection on this apparently unreflective conservatism, which has governed course design in the past, is contained in the quotation by Ashby, for many years a Vice Chancellor and a deep educational thinker who was 80 years old before he saw the possible need for change.

Traditional assessment practices, consisting pre-eminently of the assessment of essay and problem type final examinations and similarly constructed coursework, cannot adequately test for imponderables like independent critical thinking, creativity etc, and this is particularly so for time limited examinations. Although,

perhaps more so in the humanities than the sciences, good candidates can and do produce original thinking under examination pressures, such thinking under time pressures is far removed from the reflective thinking which evidences true originality and creativity. It is also more difficult in the sciences where there is a large body of knowledge on which any original thinking has to be based. This often results in questions which test first for the lower levels in the hierarchy of knowledge (recall and simple applications), not because they are more important but because they are easier to test for, and then uses that knowledge to solve problems, thereby removing most originality. That traditional assessment practices are inappropriate for many innovative approaches to teaching and learning, particularly those which place the student as central in the teaching-learning process and those which are concerned with the development of academic and/or life skills[2], has in principle been largely accepted. However, because in general it is even more difficult to assess skills reliably than content, they tend to be treated peripherally in assessment; see eg Hinett and Knight (1996).

What has been far less recognised, let alone accepted, is that traditional assessment practices, in particular the heavy reliance on timed examinations, may also be unsuitable for traditional approaches to teaching and learning, but there has been an almost total absence of research associated with the questions of Oppenheim et al.(1967) which will be discussed below. For that reason, our approach to the task in hand, when applied to traditional assessment inevitably involves a mixture of research, practice and insight.

Early innovative ventures

This is not to say that there were not some attempts at innovations even thirty years ago, eg those reported at a conference on new assessment techniques in use – in the arts, pure sciences and applied sciences (CVCP/AUT, 1969). A number of innovations were reported as being introduced into the assessment of university degrees, such as the use of oral assessment; providing more choice, eg giving students more choice in what aspects of their work each is to be assessed on and allowing different students to reveal their different strengths; allowing some assessments to be pass/fail while others are graded; project work and other forms of course work. Some of these innovations are by now standard, so that it is now difficult to remember a time when all assessment in most subjects and was on the basis of unseen written examinations (the first to introduce course work assessment were the Universities of East Anglia and Surrey in about 1967). Others are still waiting to be tried out, but it is interesting to see, how many innovations it was possible to introduce, given a will to do so, and that it could be done without offending the regulations of the institutions.

Towards an innovative traditionalism

If Oppenheim et al (1967) were and are right, there is much to be done, and I will now look at some of the assumptions listed in their paper. So far their scrutiny have – one suspects - had barely any impact on practice, either positively or negatively, and it is surely significant that in response to a recent letter to all subject LTSNs, inviting them to comment on the Oppenheim assumptions, not one suggested that any of them might be out-of-date. I selected these assumptions from the total by first making a list of problems which I considered to be important currently and then try to find an Oppenheim assumption that linked to it. I found an assumption to link, although at times somewhat tenuously, to many items on my list, although not all and not, if some logical progression was to be maintained, in the original order. Inevitably, there will be a certain lack of continuity in this chapter, but there is an underlying thread, the issues pertaining to reliability and validity, concepts which – perhaps surprisingly - did not feature explicitly in the Oppenheim assumptions.

Reliability and validity

Both reliability and validity will be treated in a more fundamental way in the next chapter, but it is appropriate to include some basic considerations here. We will be concerned with the following questions:

- Under what conditions can the same assessment be both highly reliable and valid?
- Under what conditions can either the one or the other be high?
- How concerned should one be if either is not?

[2] Academic skills have of course always been tested in context with the content in which they were exhibited, but how or even whether they or their transfer can be tested independent of context is arguable, as is whether life skills can or should be tested. The whole issue of skills is far from settled.

The two most common aspects of reliability are marker reliability (two markers give the same mark for a given assessment) and paper reliability (a given student performs equally on two supposedly equivalent papers). The results of research on marker reliability - from Hartog and Rhodes (1936) to Cox (1967b) and beyond - have always been the same, namely that the marker reliability of essay and problem style examinations was comparatively low, except where pure knowledge recall was required. Attempts to improve marker reliability relied largely on simple statistical techniques. At no time, then or now, has the professionalism of those who set or marked examinations been questioned, in spite of the fact that few if any had had any training for their task (see Ashby above). Research on paper reliability has been rare, since students are not normally subjected to two supposedly equivalent assessments under the same conditions at the same time. However, this was done by McVey (1976), who found that the best and worst students performed similarly on both, but that the rank ordering of middling students was by no means the same. The obvious explanation of this phenomenon is that middling students perform well on some and badly on other supposedly equivalent questions. The same problem of paper reliability particularly affects all except the best designed multiple choice tests, although these obviously do well on marker reliability.

The assumed basis of assessment in psychological measurement led to the prime role given to reliability, almost certainly the area where most research has been done, and in particular marker reliability, ie the extent to which different examiners agree on the mark to be given to a particular piece of work. It was appreciated that the insistence on high reliability often resulted in curriculum areas inadequately represented in examination papers, if they could be assessed only unreliably, eg any that involved any kind of originality. Such areas were then either omitted or at least under represented in comparison with their importance in the curriculum. Only recently [see eg B. Broad (2000) and M D Miller and S M Legg (1993)] has the positivist basis of assessment in psychological measurement been challenged in terms of the interpretive paradigm, with a resulting potential increase in reliability. [For a first introduction to the concepts of positivism and interpretivism see the section The organisation of this Report p.8-8.

This consideration takes us to what is called face validity, ie the extent to which an assessment is seen to assess what it is supposed to assess. Clearly the practice of under representing curriculum areas which are difficult to assess or – far worse but not uncommon – to assess what can be assessed reliably but does not relate to the declared learning objectives is wholly indefensible, particularly as it has been known since the 1970s (Becker 1968, Snyder 1971) that there is a backwash effect from assessment to the learning which precedes it, ie students take their cues as to what and how to learn far more from the assessment that they will be subjected to than from the teaching which they have received, ie low validity has seriously bad effects on learning[3]. In the real world, in contrast to that of psychometric testing, neither reliability nor validity can normally be as high as it is in such testing. There is indeed a trade off between reliability and validity and, as a consequence of a general appreciation of the backwash effect, considerations of validity have become more important over the past thirty years at the expense of those of reliability (Elton, 1982). Moss (1994) provides a more fundamental critique of the reliability v. validity issue, in terms of psychometric and interpretivist approaches to assessment.

This move started with the classic book by Rowntree (first edition 1977), who first questioned the dominance given to reliability over validity (see his outstanding account on ‘Validity and Reliability in Interpretation’ in ch. 6), and, in particular, the dominance given to ‘fairness’:

> ‘I can’t believe that ‘fairness’ in assessment should be reduced to ensuring that assessors agree. The price in terms of educational relevance is too high.’

But while Rowntree recognised that much of the unreliability of examinations was due to the lack of professionalism of marking assessors, he failed to notice that the same lack applied also to the setters of examination papers, although their work too is part of an assessor’s task. In particular, how often, in an examination where candidates are free to choose, say, any five questions out of eight, are these questions really of comparable difficulty? I do not know of any research on this subject, but it came to my attention when I was asked to write a good practice guide for University College London (see Appendix to this chapter) and looked at recent examination papers. And how many lecturers explain to their students the difference in meaning of such ‘essay words’ as discuss, compare, etc. Indeed, how many lecturers use these words consistently themselves? To make sure that students are clear about the meaning of an examination question and to be clear about it oneself are both qualities of a professional assessor.

[3] Other aspects of validity are considered in the next chapter. Strictly speaking, it will be shown from the definitions there that what we are concerned with here, namely face validity as perceived by students, is a form of consequential validity.

Descriptions of well conducted examinations exist particularly in the medical area. R M Harden (1979) gives a sound traditional overview, which is applicable quite generally, and in the Netherlands, J Cohen-Schotanus (1999) more recently concluded equally traditionally, that

- exams should be programmed regularly, ie the times between exams should be roughly constant;
- the examination rules must be fair but strict; the rule of thumb is 'the stricter the rules, the better the results';
- all educational objectives have to be assessed.

By following these rules, assessment drives student learning in a positive way.

At this general level, these prescriptions may well apply to most disciplines, but the interpretation of 'fair' and 'strict' could be very different, eg in the assessment of a critical essay in philosophy or of an artifact in art and design. In the latter, there has been a move away from the traditional 'crit' style of assessment towards one more based on predetermined intentions, as a result of which students' individual intentions tend to get ignored [H Cannatella (2001)]. This surely is a regression towards the dominance of reliability over validity.

Finally, on the subjects of reliability and validity there is the question of plagiarism, which clearly affects both. McDowell (2001) gives a useful summary of the current situation and how it is affected by mass access to higher education, changes in assessment practices, communication and information technologies and the importance of grades. She also gives good advice on the often genuine differences of perception between staff and students as to what constitutes cheating (as opposed to collaboration) and plagiarism (as opposed to deference to authority in the field of study). Academics must realise that for students the rules of plagiarism are complicated and not obvious.

Current problems – the Oppenheim assumptions

I now list some of the problems in traditional assessment which I believe are in urgent need of attention. I start with the Oppenheim assumptions, since these relate to problems perceived for more than thirty years and still unresolved. How these can or should be resolved is different for the different assumptions, some of which would appear to be rational while others are not. Whether I consider a particular assumption rational or not will, I hope, be clear from my treatment of it. I list them first as an advance organiser before treating each in some detail.

The distribution of marks

Assumption 17, that *examination results should be distributed in a certain way*.
This 'certain way' must surely be open to doubt, particularly in that it seems to be based on a fallacy – that examination performance is a form of psychometric testing.

Learning objectives and graduateness

Assumption 1, that *university examinations can include some so-called imponderables such as 'quality of mind', 'independent critical thinking', 'breadth', etc in their assessment.*
It is these 'imponderables' which it is thought are primarily what students should take away from their studies and so what distinguishes a graduate from a non-graduate.

Formative and summative assessment

Assumption 20, that *'learning is to be valued for its own sake' and not merely as preparation for a career and for financial gain.*
It is generally accepted that summative assessment strongly and often unfavourably influences a student's attitude to their learning; could formative assessment have a parallel but more beneficial effect?

The predictive aspect of assessment

Assumption 12, that *forced regurgitation of knowledge under stress is predictive of future performance*.
Employers are said to demand degree grades for their predictive value. What is the evidence for correlations between degree grades (even if based only in part on 'forced regurgitation of knowledge under stress') and subsequent performance?

Degree class

Assumption 2, that *quality of academic performance is rateable on a single continuum, from first class honours to failure*.
Any degree course has so many facets that it is intrinsically improbable that a classification in terms of a single variable can do performance in it justice.

Profiling
Assumption 21, that *the outside world wants the results of university examinations, or takes much notice of them*.
If this assumption is correct (and even if it is not), would it not be more informative and also fairer if the final report on a student's performance were in the form of a profile?

Comparisons of the standards of different degree programmes
Assumption 18, *that we are forced to make and then to retract all kinds of assumptions about the comparability of degrees from university to university and from country to country.*
Can such comparisons of entities which differ significantly and even fundamentally in many ways, be made reliably? Or are such comparisons largely fictions which are politically convenient?

Fairness
Assumption 19, that *there is a need for uniformity in undergraduate exams: all students in a given year group must pass the same examination paper, and we do not allow examinations to be tailored to individual needs.*
Is it really 'fair' to treat all students the same when they so manifestly are not? Would it not be fairer if they were allowed to show their different strengths and perhaps to hide their different weaknesses, in the way that adults normally act, once they are past being examined?
External examiners
Assumption 11*, that external examiners prevent bias.*
This is only just one of the many issues concerning external examiners. Research evidence does not seem to support the high prestige which they have in Britain. This is not to say that they should be dispensed with – they are unknown in most countries – but that the system may need reforming in the light of the evidence.

The distribution of marks

Assumption 17, that *examination results should be distributed in a certain way*. Here we have three quite separate problems:

(i) Mark distributions seem to be referred back to psychological measurement which is assumed to give normal distributions. In reality, mark distributions are often legitimately skewed, eg good teaching can reduce failure tails. Also, it is rarely appreciated how 'local' the conviction that mark distributions should be roughly normal really is, for in eg the United States grades are heavily skewed towards the positive side, so that A's are quite common and F's very rare.

(ii) The division into classes, eg over 70% for a First and 40% for a pass, in all subjects, seems completely arbitrary; the error bracket inherent in any particular mark (69% is a 2.1, but 70% is a 1st) is not understood; and the way that different legitimate conflations of marks can change their totals is not appreciated. For a devastating criticism of this last point, see the much used workshop by Fox (1982) where he states: "For some reason certain people who are otherwise quite rational become very emotionally worked up about examination marks, and they will argue fiercely, irrationally and uncompromisingly about the meanings that should be attached to such marks. . . . In particular, participants are apt to react violently to any suggestion of what they see to be 'tampering' with marks." A Year in the United States may provide a cure for such illusions, just as a year in Britain might cure some American ones. Similar points are made by Knight (1999).The subject of how different methods of conflation lead to different total marks and hence different degree classes does however seem at last to get the attention it deserves [see eg Bridges et al (2001) and Rivlin and Roberts (2001)].

(iii) Mathematical style examinations have a broader distribution of marks than essay type ones (Elton 1969). This - because of the sacrosanctness of '70% is a First, etc' and almost certainly for no other reason - leads to different proportions in a given degree class for different disciplines, eg far more Firsts as well as failures in the natural than the social sciences. (Elton 1998). Woolf and Turner (1997) showed that without greater transparency there was no way of knowing whether a given class in a given subject had the same meaning in different institutions, and Bridges et al

(1999) showed that degree class was affected by grading method.. I have recently discovered that similar problems occurred in school examinations, where they were successfully handled statistically, ie through 'marking to a curve', many years ago [Petch (1953), Nuttall (1974), J C Mathews (1985)]. However, while statistical treatments are valid for schools where numbers are large, they cannot be used for the much smaller numbers which are normal in university examinations.

Learning Objectives and graduateness

Assumption 1, that *university examinations can include some so-called imponderables such as 'quality of mind', 'independent critical thinking', 'breadth', etc in their assessment.* This is an assumption that I believe to be valid. The reason for dealing with the issue of expressing learning achievements in terms of declared objectives is that expressing them in this manner is particularly difficult for 'imponderables', like the ones listed above. The concept of expressing what students were expected to learn in terms of learning objectives came to Britain in the mid 1960s from the United States, largely as result of the work of Bloom et al (1956). [It is difficult to appreciate nearly forty years later, how revolutionary Bloom's idea – that it was possible to codify knowledge acquisition, understanding and use in terms of a hierarchy from reproduction to application, analysis, synthesis and judgment was at the time.] Bloom's taxonomy is not mentioned explicitly by Oppenheim et al (1967), but it underlies assumption 1, ie that there are learning objectives which are not strictly related to disciplines. The concept of learning objectives was first used in UK higher education by the Open University and the Council for National Academic Awards in the 1970s, and it became an important aspect of the increasing outside surveillance of universities from the mid 1980s. However, an analysis of the wording of actual questions in current finals papers reveals that, while learning objectives may figure large in submissions to the Quality Assurance Agency, it is not always apparent that they have influenced actual examinations. This is not surprising, since the common acceptance of any new idea usually takes many decades, but it does mean that much assessment practice continues to be very traditional and not in line with what might be termed 'traditional good practice' (see Appendix to this chapter).

The use of learning objectives in assessment is not straightforward. Initially, in the Open University, they tended to be close to behavioural (eg "at the end of this course a student will be able to list five important features of . . ."), and there is the Open University story of the External Examiner from a traditional university who, when told that these objectives were revealed to the students, exclaimed: "But in that case they will all pass!". What this approach to objectives fails to grasp is that any worthwhile learning has elements of uncertainty, which makes pre-specification in terms of deterministic behavioural or near-behavioural objectives impossible. At the same time, this is not a justification for leaving the objectives tacit or not to have them at all, for learning objectives are an essential aspect of good teaching and learning. However, to formulate such objectives so that they are informative and yet not constraining is often an art and never a science.

Once learning objectives are specified in a non-behavioural or at least less than deterministic manner, judgment enters into the extent to which a particular student has achieved them. This judgment used to depend on the tacit knowledge of experienced examiners and, although in many instances this knowledge has now been codified so that it is no longer tacit - eg definitions about different operational essay words such as 'discuss', 'compare', etc (see Appendix to this chapter) - the judgment still depends on the experience of examiners who are almost uniformly untrained (see below). A rare glimpse at the views of students is provided by Powers and Fowles (1999). The study identified several features that underlie examinee perceptions – often unsound - of essay prompts, eg the extent to which prompts allowed examinees to draw on their own personal experiences. A more general view of both staff and student reactions can be obtained from Maclellan (2001). Overall, there was little comfort to be obtained: 'staff maintained that the full range of learning was frequently assessed, yet the dominant mode of assessment was the traditional academic essay', ie they deceived themselves, while 'students do not exploit assessment to improve their learning', ie they missed out on a most important way that examinations could help them..

Assumption 1 itself, ie that examinations can include some so-called imponderables, has led to the development of detailed learning outcomes**,** eg in National Vocational Qualifications, in which these general imponderables are broken down into elements and assumed - contrary to all experience – to be no more than the sum of their parts. The difficulties inherent in such an outcomes approach to assessment were well rehearsed by Otter (1995), who nevertheless – and to some extent against her own evidence - was a proponent of it. In fact, some of these imponderables, such as 'independent critical thinking' do not lead at all to learning outcomes, but can only be described in terms of processes. A holistic approach to assessment through what is called connoisseurship [Eisner, 1985)], which can deal both with both processes and products, will be discussed below.

A solution to the problem of assessing imponderables became a major task for the Higher Education Quality Council, a body independent of Government, in its search in the 1990s for criteria of graduateness, since any specification as to what distinguishes a graduate from a non-graduate in their learning achievements is likely to depend on such discipline independent imponderables. Their conclusion was [HEQC (1996), sections 8.4 and 9.2]:

> 8.4. Overall it is felt that the assessment procedure in higher education is not compatible with a threshold model in the sense of a sector-wide set of outcomes that would define explicitly and meaningfully what either a particular sort of graduate, or any sort of graduate, 'is'.
>
> 9.2 A common vocabulary was in use across disciplines and across institutions, but cannot be taken in itself to imply much commonality in standards.
> . . .
> Most current assessors were socialised into a system in which the links between course creation, assessment and awarding were very close. The effects of current trends towards fragmented marking, formula-driven awards and small examination boards are not currently being offset by any other developments. [HEQC was too kind to the past here. Ever since Hartog and Rhodes (1936) it has been known that such socialisation – which indeed led to the 'tacit knowledge of experienced but untrained examiners' was an inadequate guarantee of reliability. LE]
> . . .
> Guidance and criteria relating to outcomes and levels are often phrased at a very general level and do not appear to be used much at the point of assessment.

However, this did not mean that nothing could be done. HEQC argued that, whatever may have been the situation in the past, the future required greater attention to be paid to the course approval process and training for staff and its major recommendation related to the development of professional and subject body involvement (see HEQC sections 9.8 – 9.14.) Such measures were designed to make examiners more professional, in line with the likelihood that the single greatest uncertainty in assessment arises from the lack of professionalism of examiners. A good exampleof this uncertainty is provided by Nelson (1990), where "students interpreted writing assignments in a variety of courses and how these interpretations differed from their teachers' intentions." Nelson argued that these misunderstandings arose from the complexities of communication between students and their teachers, who were also their examiners, but the understanding of such complexities is surely of the very essence of professionalism of teachers as examiners. However, the idea that professionalism will enable marker agreement to be near perfect is wrong; there are disagreements even between professionals [see eg Huot (1996)] which will be discussed further in the next chapter].

While the search for criteria for a standard of graduateness in general proved unprofitable, it might be possible to establish such standards for individual disciplines.
This idea was taken up by the Quality Assurance Agency in their benchmarking programme, which tries to specify standards at different levels for individual disciplines. An interesting aspect of this programme is the frequent use of words which require interpretive judgment. Thus in Chemistry, the top three levels of conceptional understanding are characterised as: outstanding, good and generally sound, while the lowest level - still acceptable for a pass at degree level! - is devoid of conceptual understanding. In Education, the three acceptable levels of 'different contexts in which learning can take place' (to give just one example) are respectively: a well developed understanding, a good understanding, an awareness. Such phrases can certainly help to inform judgment; they do not replace it.

The concept of informed judgment forms the basis of Eisner's approach to the assessment of communicative and interpretive understanding, and of self-reflective and critical knowledge. It also forms the basis of the approach of Toohey (1999), pp. 175 – 180, to assessment. Eisner uses the terms 'educational connoisseurship' and 'educational criticism', to characterise the ways that professional assessors perform tasks similar to those of connoisseurs and critics in the arts. Such a totally holistic approach to assessment is certainly possible for much educational judgment, but it must go beyond the traditional 'I can recognise a 2.1 when I see one'. Such connoisseurs in fact, if only tacitly, must make their judgments in a more detailed manner which, at least in education, should not remain tacit. For that reason perhaps a better model (Elton 1998) is that of the judgment of the performance in an Olympic figure skating contest, which is judged holistically through guidelines on a number of dimensions expressed in qualitative statements. These are then turned into numerical grades and combined with appropriate

weights into an overall assessment. [Brown et al (1997), pp. 128 – 129, give good examples of Guidelines in terms of dimensions for projects and for dissertations.]

The analogy with Olympic judges may indeed provide an opportunity for referring to a concept - that of an 'interpretive community' of assessors - which is treated much more fully in the next chapter[4]. The decision making of such a community is very close to that of peer judgment of, eg papers for publication or RAE grades, but it is only valid if the peers are experienced and knowledgeable in the area to be judged[5]. It also relates to Eisner's connoisseurship, particularly when modified, as I have suggested above, to the model of Olympic judges of figure skating, where judges make individual judgments on the basis of all being members of the same interpretive community. The parallel can be taken further: their judgment of the Compulsory part is more positivist, ie on the basis of a pre-determined schedule, while the judgment of the Voluntary part of the performance is closer to being interpretivist, in that the criteria on which the assessments are to be based are agreed, but there is no assessment schedule. On the other hand, it is not fully interpretivist, as that would probably involve the criteria being established in the light of the examinees' performances through discussion between assessors. However, in combining scores at any time, it is important to remember that "the practice of combining scores is so entrenched that it would come as a great surprise to many to be told that it does not necessarily have to be done this way and that there are convincing reasons for abandoning the practice."[R Wood (1991)].

Examples of such qualitative statements against which performance can be judged and which relate to higher education are given eg in HEQC (1996), App. A. However, Ecclestone (2001) argues that such statements are inadequate, that there is a need for a common interpretation of such criteria among all in a relevant academic community, and that this requires a more strategic approach to inducting and socialising staff into that community. So here again, there appears to be a need for an interpretive community, which in the first place will have to decide in each case whether to approach its task in a positivist or interpretivist manner. This will certainly involve them in discussions of reliability and validity.

One may be able to detect a parallel between the intellectual development of the assessment community towards more sophisticated approaches to assessment and the well known work by Perry (1970) on the intellectual development of students, however far fetched this may appear at first glance. Positivism would then correspond to the certainty of the lower levels of the Perry hierarchy; the views, held largely tacitly of experienced but untrained examiners representing the 'anything goes' middle stage; while interpretivism corresponds to the upper levels of commitment, with BJ in chapter 3 firmly warning against any suggestion of interpretivism representing the 'anything goes' middle stage. However, it is at least arguable that at the extreme of the upper levels, commitment ought to be exhibited not only by the examiners but also the candidates, with the latter choosing their own commitment stance.[6]

Formative and summative assessment

Assumption 20, that *'learning is to be valued for its own sake' and not merely as preparation for a career and for financial gain.* This hope relates strongly to the different purposes of summative and formative assessment. The difference between the two purposes of assessment, ie summative for judgement and formative for improvement, characterised by Knight (2001) rather nicely as 'feedout' and feedback', has been well known for a long time. The problem with formative assessment has always been that it is essential for good learning, but that students may not take it seriously, as it does not 'count'. Thus I have heard it said by Open University tutors that they frequently find their students conscientious about the tutor marked assignments, but not about the earlier assignments which are there for practice and improvement. One way to overcome this problem is to mark an assignment twice, a first deadline for improvement with feed back,

[4] See *Interpretation and the interpretive community* p.44-45.
[5] The use of peers, who are eminent and knowledgeable in research but have little beyond their own experience in teaching, learning and assessment, as assessors in the latter is not unknown.
[6] BJ points out that there is an essential problem in using the work of Perry, which is firmly in the positivist frame, for anything that involves also the interpretivist frame, but as a thought provoking comparison it may be acceptable.

and a second for judgment. It is then up to the students whether they want to take advantage of the possibility of feedback or not. Such a scheme can become quite sophisticated and flexible, eg there might or might not be a tutorial between the two submissions, there might be additional library searches, it could become a staged assessment rather than a twice marked assessment, etc. Here Knight's paper has challenged the established orthodoxy that - since formative assessment thrives on students' openness, while summative discourages it - the same assessment cannot be used both formatively and summatively,. Going beyond the two-part assessment referred to above, he suggests that, minimally, summative course work assessment should provide feedback.

Tackling the issue from a different angle Knight (2002), ch 9, suggests that summative assessment has to be as precise as possible, which makes it expensive, but that this is not so for formative assessment. This makes it possible to use less costly and possibly less reliable methods, including self and peer assessment. One can then go on from there to the assessment of skills, such as group skills where self and peer assessment may be the only possible form even summatively, and mental skills, such as criticality, where the development of the skill and its assessment cannot sensibly be divorced from each other. The drift of Knight's argument is that if students are given a real stake in their own learning, they will learn better, with more enthusiasm and with less of an eye on their summative assessment when such assessment is needed, which is not always the case. A highly relevant review of formative assessment in schools has been provided by Black and Wiliam (1998), who also affirm Brown's maxim, quoted earlier, that 'if you want to change student learning, change the assessment' which incidentally was formulated in those very words already twenty years earlier [Elton and Laurillard (1979), p. 100.] but forgotten. It may well not have been new in 1979![7]

Different perceptions of staff and students regarding formative assessment have recently been studied by Maclellan (2001). She found that 'while staff declared a commitment to the formative purposes of assessment . . .and maintained that the full range of learning was frequently assessed, they engaged in practices which militated against formative assessment . . . being fully realised.' Students, on the other hand, did not 'exploit assessment to improve their learning'.

But what has all this to do with the hope of academics that students should learn for its own sake? The link is that formative assessment should not confine itself to what will eventually be summatively assessed, but rise above it. For this to happen, students have to become interested in what they are studying for its own sake and they are normally only prepared to do that, if their needs to be prepared for the summative assessment is first satisfied. This need overrides everything else, in line with the well known psychological model of Maslow's hierarchy of needs, ie lower needs have to be satisfied before the higher ones will be considered, a hierarchy graphically expressed by Brecht in his aphorism 'First fill your stomach, then come morals'. However once the summative assessment needs have been met through appropriate formative assessment, students are willing to consider higher aims such as learning for its own sake (Elton 1995**)** and further formative assessment can then be linked to such learning.

The predictive aspect of assessment

Assumption 12, that *forced regurgitation of knowledge under stress is predictive of future performance*. This assumption makes the claim that a major, although not of course the only component of many, if not all examinations, consists of assessing regurgitation of knowledge under stress. Few students would deny this. Now, Hudson (1960) researched into the predictive power of examinations in a specially favourable case, the later achievement of Oxbridge science graduates who became Fellows of the Royal Society or obtained an Oxbridge DSc. His conclusion was that "faith in Oxford and Cambridge degree class as a predictive index of potential research ability in science should be viewed with scepticism". He followed this up with a book, Hudson (1967), in which he concluded (see ch. 6) that "divergers are potentially creative, convergers are not", so that creativity tests – if they can be made reliable, which may be an impossible task – would be a better predictor of research success than degree class.

Hudson could carry out his investigation, because thirty five years ago a major qualification for becoming a PhD student was to be able to afford it and many with Third class degrees were taken on for research. Such an investigation could not be repeated now, since in research as well as in most other first employment situations, the predictive value of the degree class is assumed to be valid, so that those with poor degrees are disadvantaged. Recent evidence by Blasko (2001) indicates that those with better degrees

[7] Not for one moment would I suggest plagiarism, but it does illustrate the problems associated with that concept.

certainly tend to get higher salaries, but it also indicates a selection process which may make this prediction self-fulfilling. It may be worth adding an anecdote here. I have a memory that many years ago ICI had a more sophisticated recruitment policy. It preferred introverts with first class degrees for research and extroverts with third class degrees for the sales force. Some years later, the latter had higher salaries than the former!

Current interpretations of using Graduate Employment as a performance indicator for higher education have been challenged as simplistic by Little (2001). Furthermore, as degree performance is much more related to knowledge than to skills and skills can rarely be assessed as reliably as knowledge, the increasing importance attached by employers to the acquisition of skills may well make degree results even less relevant predictively than has been the case in the past. It is perhaps fortunately not relevant to the purposes of this project whether the increasing demands by employers for the inclusion of a variety of skills in undergraduate curricula can be justified on academic and pedagogic grounds.

Having said all this, I must return to the formulation of Assumption 12 which characterised traditional final examination as '*forced regurgitation of knowledge under stress'.* How far final examinations are mainly reproduction of previous learning is very difficult to assess, but in one subject it was indeed found to be so (Black, 1967). It is actually often very difficult for an external examiner to distinguish between original student's thought in the examination and teacher's original thought in lectures reproduced in the examination, but at least one former student (BJ, private communication) claims that

> 'if you regurgitated, you got low grades; you had to adapt and think how what you did know about a subject could be used to address the question you had been set. That required a lot of initiative and creativity and fast thinking.'

Degree class

Assumption 2, that *quality of academic performance is rateable on a single continuum, from first class honours to failure*. This assumption is so patently absurd – we grade even eggs on two scales, size and age - that when universities were treated in the same way by the Quality Assurance Agency, there was an outcry and the single scale was replaced by six scales which tested different aspects of performance. This did not of course stop the six scales being meaninglessly conflated into one and reported in that way in league tables. The problem has become immeasurably aggravated in the past thirty years, as the monolithic finals exams have become multidisciplinary in modular degrees and the traditional written papers have been supplemented by a greater variety of forms of assessment – course work, work placements, etc - which rarely test for the same learning objectives. One way to overcome this problem of conflation of marks is to abandon degree classifications and replace degree certificates by profiles which relate to different types of achievement [see Fenwick et al (1992)]. At least some of these should not however be on numerical scales, because if all are, whether it is appropriate or not and it rarely if ever is for all, then they are inevitably averaged - as is the case in the American transcript - into a 'grade point average', which becomes the only thing that is really considered to matter and is very similar to the British degree class. The recognition that not all achievements can be numerically graded is important, but it has to be admitted that there are occasions when an overall quantitative grade may be required as a result of qualitative judgment, as was suggested above in connection with Olympic figure skating competitions. However, I would suggest that such grading should then be carried out by the user at the point of use; in the Olympics to identify the gold medallist, in degree statements in connection with first employment, research potential etc. In fact, a profile might well contain items which simply cannot be meaningfully related to a numerical scale, such as endeavour or creativity. [I have to confess that when I first introduced profiles at Surrey (Elton 1969), I got no further than to concede that the assessment of such qualities 'bristles with difficulties', but I believe that many of these difficulties have by now been tackled.]

While I believe that the educational arguments against overall final grades are now overwhelming, there is the argument that a final grade is demanded by employers, when they select candidates for employment. They cannot do this in countries (and there are many) where degrees are not classified, and even in this country it is probably rare for degree class to be used except in preliminary sifts of job applicants. To find acceptable ways for preliminary sifting is important and, if the advent of profiles led to employers making preliminary sifts on the basis of A level performance, that would be serious. Clearly, a discussion between universities and employers is indicated.

Profiling

Assumption 21, that *the outside world wants the results of university examinations, or takes much notice of them*. One of the most serious obstacles to good and meaningful summative assessment is the requirement that all separate assessments must in the end be summed into one overall assessment, the degree class. This absurdity was already challenged above in connection with the Oppenheim assumption 2 and I now want to take this matter further than I did there.

The kind of soft assessment, eg of mental and group skills, cannot be put in a strait jacket of grades [see eg Edwards and Knight (eds)(1995), pp. 10 – 24, who argue strongly for a profile approach and document the inadequacy of traditional approaches]; indeed in many instances eg the development of leadership skills, team skills, skills of empathy etc, it is only the process that can be reported and not the product. Hence, if a degree curriculum involves, as it surely does, the development of certain skills, then the choice for employers would appear to be between the misleading accuracy of a simple grade and a much more detailed statement which has to be matched against specific requirements.

A subset of these skills which is of particular interest to employers is that of transferable skills, ie skills which transfer from the educational experience into employment, and it has indeed been suggested that there are general skills of transfer. The subject has been discussed widely. Fleming (1991) defines them as "the metaskills, the second-order skills which enable one to select, adapt, adjust and apply one's skills to different situations, across different social contexts and . . .across different cognitive domains". Bridges (1993) in an important article makes the distinction between transferable skills and the skills of transfer. The former are skills which, when learned in one context, can be transferred and used in another. The latter is the skill which enables one to make such a transfer. Whether such a metaskill actually exists is by no means agreed; the question is raised by Bridges (ed), 1994. Assiter (ed) (1995) in an otherwise very informative book never refers to their assessment and Becher (1994) argues that "many of the study skills programmes for students are general in nature and seen as of limited use by their participants. . . The whole mode of argumentation differs radically between such fields as biochemistry, English literature and the sociology of science", in other words, participants did not see them as transferring from the general to their particular needs. This confirms similar points made repeatedly, particularly in connection with graduates in the humanities in eg Eggins (ed) (1992). Against that, Alverno College (1984) claims that there is a set of skills

1. observe accurately
2. make justifiable inferences
3. relate parts or elements in patterns
4. integrate patterns into whole systems
5. compare and test frameworks in own discipline
6. integrate frameworks into a professional synthesis

of which the first four are generic and only the last two subject specific. Hinett (1995), gives a perceptive analysis of the different approaches to assessment by Alverno College, where assessment is formative and an integral part of learning, and by a typical British university, where it is summative and linked to external accountability, much to the detriment of learning. If Alverno College is right, then the reason why students find it so difficult to transfer a skill is that they have learned it in the wrong way, as have their teachers before them. See also Assiter (1993). Finally, in this discussion, I would like to quote another passage by Ashby (1963), talking about his fellow academics:

> 'All over the country these groups of scholars, who would not make a decision about the shape of a leaf or the derivation of a word or the author of a manuscript without painstakingly assembling the evidence, make decisions about admission policy, size of universities, staff-student ratios, content of courses and similar issues, based on dubious assumptions, scrappy data and mere hunch.'

If Ashby is right - and few who have for some years attended university committees are likely to deny that he is - then academics rarely transfer the attitudes, which they normally have towards their research, to other of their activities. And if they find the skill of transfer difficult, how can they develop it in their students?

Although it is generally accepted that for skills it is the process of their development that is so important, a process itself is not observable. What can be observed is a series of successive products, and it is these which can be assessed. Nevertheless, the assessment of skills is essentially different from the assessment of knowledge, and this strengthens the case for reporting a student's degree experience in terms of a detailed profile, and not in terms of a single grade. It should be stressed that such a profile is simply a report of the student's learning experience; while it may be compiled entirely by the student's teachers, a strong case can be made for including in it matters chosen by the student. It may, but does not need to contain a student's personal development profile, where the word 'profile' is used in a very different sense [Heywood (2000), pp. 312 – 315]. Such a personal development profile, also known as a 'Record of Achievement' [see

Assiter and Shaw (eds) (1993)] could be an excellent way for students to demonstrate their achievements, particularly if it is used formatively. In practice it faces formidable difficulties. If it does not form part of the overall assessment, it tends not to be taken seriously; if it does, it creates tensions between formative and summative assessment. An excellent critical account, which includes a change theory analysis of the adoption of such an innovation into a conservative system, has been given by Trowler and Hinett (1994).

One huge advantage of profiling over degree grades is that the pressure to make assessments reliable can be concentrated on areas where it is possible, which can then be numerically reported, while others are reported descriptively. In a conference on profiling (Winter 1994) the only argument against profiling that made even remotely sense was that employers would find it more difficult to select future employees, or rather provide them with an easy first sift, because no rational employer – other than apparently the Research Councils! - would use a degree class as an important selection component for an actual appointment. This argument has to be set not only against all the arguments that favour profiling, but also – as was argued above - against the argument that in many countries degrees are not classified in the way that they are in Britain. Finally there is the argument that reporting in terms of profiles might greatly increase staff workload, although it is not clear without a serious pilot study whether this is really so.

In this connection, the question whether 'the outside world wants the results of university examinations, or takes much notice of them' is not easy to answer. The outside world has changed radically in the past thirty years and while in general it is probably still pretty relaxed about degree results, Government has made them an important performance indicator. Whether it should do this is, to say the least, arguable (Pollitt 1987). There is also the possibility that the outside world might be satisfied by a mere statement that a degree has been obtained, and there has been a movement for the abolition of the classified degree (Winter 1993).

But the biggest argument for going over to a profile is educational. Whatever may have been the argument for a classified degree when the university system was uniform and all assessment was based on performance in finals papers, the present huge variety of degree offerings and forms of assessment cannot be served well by such a system. Under such conditions, the demand for an overall degree class distorts all assessment and the degree class itself loses meaning. At the very least, the time has come to look at the problem seriously and dispassionately. It is 30 years since Powell and Butterworth (see p.6) first raised it.

Fairness

Assumption 19, that *there is a need for uniformity in undergraduate exams: all students in a given year group must pass the same examination paper, and we do not allow examinations to be tailored to individual needs.* The argument that only by treating all students identically can assessment be fair would have had the approval of the mythological Procrustes, who had a standard size of bed for all his guests and stretched those who were too short, while chopping off the legs of those who were too long. It can be countered by the argument that we are all different from each other and to treat us as if we are all the same is not only extremely unfair in itself, but unintentionally disadvantages those who happen to be different from the norm favoured by the assessment. Why should it be left to chance whether particular students can demonstrate their best, rather than give them the deliberate opportunity to do so?

Comparisons of the standards of different degree programmes

Assumption 18, *that we are forced to make and then to retract all kinds of assumptions about the comparability of degrees from university to university and from country to country.* To this assumption there should be added the problem of comparing degree programmes in different disciplines in the same institution, a problem that is basic to all modular degree programmes, but which has proved totally intractable. There is just no way that the standard of a degree in, say, engineering can be compared with one in, say, philosophy in terms of objectives, content and methods. In consequence, comparisons have been made in terms of credit equivalences, where credits are measured in terms of notional study time and level of difficulty. This is of course possible, whether it is meaningful is difficult to say. The difficulty and indeed impossibility of comparing degree standards in different institutions has recently been re-affirmed by Murphy (1995). Nevertheless, over the past years, as universities have diversified, there has been increasing pressure from outside on comparability of standards – between degrees in different subjects within one university, between degrees in different universities etc, pressures which could drive institutions and curricula into most undesirable conformity. At the same time, many differences are essentially historical and difficult to justify. To bring total order into this chaos is impossible, in order to bring some order where it does not harm, legitimate diversity is necessary. The problem is now a European one and the Bologna agreement is attempting to achieve such order for the whole of Europe. Its approach in terms of credit ratings at a number of levels, with the credits determined for each programme by the institution responsible

for it, runs the severe danger of creating an illusion of conformity within a reality of undetected and undetectable diversity. The jury is out on this point.

External examiners

Assumption11*, that external examiners prevent bias.* There is extraordinarily little research evidence on the effectiveness of any of the work of external examiners, but what there is does not inspire confidence in the system. A paper by Williams (1979) is based solely on his personal experience. D Warren Piper is undoubtedly the most important researcher in the field. He first issued a questionnaire (Warren Piper 1985), which concluded that

- Around half the external examiners saw all the scripts and not just a sample;
- Only two thirds regarded ensuring comparable standards between institutions as a prominent duty;
- All polytechnic but only a quarter of university external examiners committed their comments to paper.

He was then commissioned by the Department of Education to conduct a substantial enquiry and although he submitted his report, it was never published. A weaker version eventually saw the light of day in Warren Piper (1988), in which he wrote:

> "In the course of a recent research project I had occasion to interview a number of senior academics on the matter of undergraduate examinations and to analyse a large number of questionnaires from examiners on undergraduate courses. I think that without exception the respondents were people of notable intellectual sophistication and I am sure that they had all thought deeply about their subject and its examination. They shared a strong ethic about examination standards and fairness to candidates. What they did not share was a knowledge of, for instance, the compelling evidence which has accumulated over the last 20 years or so that university examinations often fail to test the kind of learning which examiners intend. Some examiners had a poor grasp of quite simple statistical theory pertinent to the aggregation of examination marks. The decisions which they made as examiners were informed by experience rather than by theory or systematically collected evidence. In many cases, the experience which examiner had to call upon was extensive, but not always." (p. 241)

Some of his findings appeared in D Warren Piper (1994). His main conclusion was that they were not, largely because they had received no training or development for their task as examiners. Elton (1989) concluded – perhaps too optimistically – that a single change to the present system, namely the pedagogic training of examiners and of university teachers in general, would result in current criticisms becoming a thing of the past. Twelve years later it would appear that this proposal is about to be accepted by HEFCE and DES in their new approach to quality assurance, in which they ask that "Universities UK should take the lead in pressing for proper training and better pay for external examiners" [Baty (2001)]. There is also a discussion paper of the Council for National Academic Awards (CNAA 1992) which drew attention to the 'rapidly growing awareness of difficulties being faced by external examiners as the structures of higher education courses change'. It ends with sixteen questions, of which one does relate to examiner training, but there were clearly many others. The last was: 'Can the external examiner system survive in its present form, or in any form?' Finally, there is a recent suggestion by Biggs (2001), that the function of the external examiners should be consultative, on the well established grounds that it is rare for them totally to comprehend the context of teaching associated with the assessment which they are adjudicating. Traditionally, and against what evidence there is, external examiners are considered to be an important and valuable feature of British universities, and they feature as the main guarantor of quality in the response by University Vice-Chancellors to the recent Consultation Paper on quality assurance. They cannot have read the sixteen CNAA questions nor the outcomes of Warren Piper's research.

Current problems – not covered in the Oppenheim assumptions

The fact that new and important topics have arisen in the past thirty five years may be a sign of progress. It is certainly interesting to reflect on why these were not considered important at the time.

Management of Assessment

Universities have a long record of amateur management. Are they justified in continuing that practice, particularly in the light of the next issue?

Assessing with reduced resources

Happy 1967 where it may not have been necessary to worry about reduced resources.

Management of Assessment

The management of assessment has, according to Yorke (1998), received surprisingly little attention in the literature. He provides a checklist of points to be covered by management, relating to:

- a functional analysis of the institutional system for assessment
- an audit of the system regarding its strengths and weaknesses
- the need to give assessment a higher profile within curriculum design and implementation
- continuing staff development in the light of the functional analysis
- ongoing consideration to the conceptual, structural and temporal relationships of assessment to the curriculum.

If these points do nothing else, they highlight the abiding amateurishness of the way that assessment continues to be treated in most universities.

Assessing with reduced resources

A consideration outside the mind set of Oppenheim et al (1967) in happier days was that of cost effectiveness. Knight (2000) argues that since reliability is important to stakeholders but difficult to achieve cost effectively, a systemic approach should be used so that resources are freed to invest in securing more reliable assessments where they are desirable and might reasonably be had. Useful advice on cost effectiveness in general is given by Andresen et al (1989).

The other Twelve Assumptions

I have referred to only nine of the twenty one assumptions of Oppenheim et al (1967). This is not to say that the other twelve are not important or that everyone would agree that this article has picked out the most important ones. So, in the spirit of reader driven learning, here are the others, with brief explanations derived from the original paper. Please comment on any that you consider important and add others. The outcomes of your considerations will be gratefully received by me at l.elton@pcps.ucl.ac.uk.

What do exams measure and how?

3. *Assessing practical work and work experience*
   Is that actually easier to assess than theory?
4. *Examinations are a mock-real-life performance.*
   Hence relevant life skills are assessed by proxy.
5. *Examinees have individual responsibility and there is no collaboration.*
   This contradicts 'real life'.
6. *A student who fails has only himself to blame.*
   Could there be something wrong with the teaching or system?
7. *Proper place for exams is at the end of a course.*
   Is there a case for interim exams, to give a better indication of future performance, or later ones, which allow for reflection?

Who should assess and how?

8. *University teachers should also be examiners and selectors*
   By giving students numbers, teachers want to distance themselves, but would not allow outsiders to do the examining. It also makes it impossible to use any other relevant information.
9. *Impartiality of examinations*
   Exclude those from examining who know an examinee best.
10. *University is sole authority to examine.*
    Neither parents nor future employers are allowed to be involved.

Psychological assumptions

13. *Mental growth and development influence the type and timing of exams.*
    We seem to base our thinking about growth and development on the following analogies:
    - Culinary: separate courses are like the ingredients in a recipe – some form of cooking is necessary.
    - Horticultural: We talk about maturing, ripening, bearing fruit, etc; merely mixing ingredients or filling casks is not education.

- Packing station analogies –'gate keeping' functions of examinations (rather like grading eggs), standards and classes.

[To savour(!) the discussion of this matter fully one has to read this passage in full.]

14. *Pressure is required.*
    Fixed examination times (in some continental countries, students present themselves when they are ready), narrow gates, peer pressures, pressure of 'having to know everything at once' will integrate what teachers have failed to integrate in their teaching.
15. *Anxiety is necessary.*
    Examination as initiation ceremony, character building etc. Mental breakdown is student's 'fault'.
16. *Quality of mind.*
    Psychological issues not covered in 1. above.

References

Alverno College (1984), Analysis and Communication at Alverno: an approach to critical thinking, Wisconsin: Alverno Productions.

L. Andresen, P. Nightingale, D. Boud and D. Magin (1989), *Strategies for assessing students,* Professional Development Centre, University of New South Wales; reprinted by the Educational Methods Unit, Oxford Polytechnic.

E. Ashby (1963), Decision making in the academic world, in Halmos, P (ed), *Sociological Studies in British University Education*, University of Keele.

E Ashby (1984), Foreword to I M Brewer, *Learning more and teaching less*, Society for Research into Higher Education & NFER-Nelson.

A. Assiter (1993), Skills and knowledge, *Reflections on Higher Education* 5 (July) 110 – 124.

A. Assiter (ed) (1995) *Transferable skills in higher education*, London: Kogan Page,

A. Assiter and E. Shaw (eds) (1993), *Using records of achievement in higher education*, London: Kogan Page.

M. J. Atkins, J. Beattie and W. B. Dockerell, 1993, *Assessment Issues in Higher Education*, Department of Employment, p.26 – 27

P. Baty (2001), *Light touch to get hard edge*, THES 30 November, p. 7.

T. Becher (1994), The Significance of Disciplinary Differences, *Studies in Higher Education* 19, 153 – 163.

H. S. Becker et al (1968), *Making the Grade: the Academic Side of College Life*, New York: Wiley.

J Biggs (2001), The reflective institution: assuring and enhancing the quality of learning, *Higher Education* 41, 221 – 238

P. Black (1968), University Examination*s*, *Physics Education* 4, pp. 93 – 99.

P. Black and D. Wiliam (1998), Assessment and Classroom Learning, *Assessment in Education* 5, pp. 7 – 74

Z. Blasko (2001), *Graduates from disadvantaged social backgrounds in the labour market*, London: Centre for Higher Education and Information, Paper presented at the CHER conference, Dijon, September 2001.

B. S. Bloom et al (1956), *Taxonomy of Educational Objectives*, London: Longmans.

D. Bridges (1993), Transferable skills: a philosophical approac*h*, *Studies in Higher Education* 18, 43 –51.

D Bridges (ed), (1994), *Transferable skills in higher education*, UEA Press, Norwich.

P. Bridges et al (2001), Coursework Marks High, Examination Marks Low: discuss, *Assessment & Evaluation in Higher Education* 27(1), 35 – 48.

P. Bridges et al 1999, *Discipline related marking behaviour using percentage*, Assessment and Evaluation in Higher Education 24, 285 – 300.

B. Broad (2000), *Pulling your hair out: Crises of Standardization in Communal Writing Assessment*, Research in the Teaching of English 35, 213 –260.

H. Cannatella (2001), Art Assessment, *Assessment and Evaluation in Higher Education* 26, 319 – 326.

CNAA (1992), *The External Examiner and Curriculum Change*, Discussion Paper 7, London: Council for National Academic Awards.

J Cohen-Schotanus (1999), Student assessment and examination rules, *Medical Teacher* 21, 318 – 321.

R. Cox (1967a), Resistance to change in examining, *Universities Quarterly* 21, 352 - 358.

R. Cox (1967b), Examinations and Higher Education, *Universities Quarterly* 21, 292 –340

CVCP/AUT 1969, Assessment of undergraduate performance, London: Committee of Vice-Chancellors and Principals.

K. Ecclestone (2001), I know a 2.1 when I see it': understanding criteria for degree classifications in franchised university programmes, Journal of Further and Higher Education 25, 301 – 314.

A. Edwards and P. Knight (eds)(1995), *Assessing Competence*, London: Kogan Page.

H. Eggins (ed) (1992), *Arts Graduates, their Skills and their Employment*, London: Falmer Press.

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

L. Elton (1969), *New Assesment Techniques – the pure sciences*, in *Assessment of undergraduate performance,* London: Committee of Vice-Chancellors and Principals, pp. 26 –33.
L. Elton (1982), *Assessment for Learning,* in D. Bligh (ed), *Professionalism and Flexibility in Learning,* Society for Research into Higher Education, pp. 106 - 135.
L. Elton (1989) *Assessment and the External Examiner System*, International Seminar on 'Assessing quality in higher education', Cambridge.
L. Elton (1995), *Strategies to enhance student motivation: a conceptual analysis, Studies in Higher Education* 21, 57 –68.
L. Elton (1998) Are UK degree standards going up, down or sideways?, *Studies in Higher Education* 23, 35 – 42.
L. Elton and D.M. Laurillard (1979), Trends in Research on Student Learning, *Studies in Higher Education* 4, 87 – 102.
A. Fenwick, A. Assiter and N. Nixon (1992), *Profiling in Higher Education*, London: Council for National Academic Awards.
F. D. Fleming (1991), The concept of meta-competence, *Competence and Assessment* 16, 10.
D Fox*, 'What gets the marks?'*, in P Cryer, ed (1982), Training Activities for Teachers in Higher Education 1, 95 – 121, Society for Research into Higher Education.
R M Harden (1979), Assess students: An Overview, *Medical Teacher* 1, 65 – 70.
P. Hartog and E. C. Rhodes (1936), *The Marks of Examiners*, London: Macmillan.
HEQC (1996), *Graduate Standard Programme*, Higher Education Quality Council.
K. Hinett (1995), Fighting the assessment war: the idea of assessment-in-learning, *Quality in Higher Education* 1, pp.211 – 222.
Hinett and P Knight (1996), Quality and Assessment, *Quality Assurance in Higher Education* 4(3), pp. 3 – 10.
L. Hudson 1960 Degree class and attainment in scientific research, *British Journal of Psychology* 51, 67 – 73.
L. Hudson, 1967, *Contrary Imaginations*, Harmondsworth: Pelican Books
B Huot (1996), Towards a new theory of writing assessment, *CCC* 47, pp. 549 – 566
P. Knight (1999), Get the assessment right and everything else will follow, *Quality in Higher Education* 5, 101 – 105 .
P. Knight (2001), *A Briefing on Concepts: Formative and summative, criterion and norm-referenced assessment,* York: LTSN Generic Centre.
P. Knight (2002), *Being a Teacher in Higher Education*, Society for Research into Higher Education and Open University Press, ch. 9. In the press.
B. Little (2001), Reading between the lines of graduate employment, *Quality in Higher Education* 7, 121 - 129.
E. Maclellan (2001), Assessment for Learning: the differing perceptions of tutors and students, *Assessment and Evaluation in Higher Education* 26, 307 – 318.
J. C. Mathews (1985), *Examinations: a commentary*, London: Allen and Unwin
L. McDowell (2001), *Assessing students: Cheating and Plagiarism,* York: Institute for Learning and Teaching .
P. J. McVey (1976), The 'paper error' of two examinations in electrical engineering, *Physics Education* January, pp. 58 – 60.
M. D. Miller and S. M. Legg (1993), Alternative Assessment in a High-Stakes Environment, *Educational Measurement: Issues and Practice*, Summer, pp. 9 – 15.
P. M. Moss (1994) Can there be validity without reliability? *Educational Research* 23(2), 5 – 12.
R. Murphy (1995), Firsts among equals, the case of British university degrees, *British Journal of Curriculum and Assessment* 5, 38 – 45.
J. Nelson (1990), This was an easy assignment: examining how students interpret academic writing tasks, *Research in the Teaching of English*, 24, 362 – 396.
D. L. Nuttall (1974), *Comparability of standards between subjects*, Schools Council Examinations Bulletin Nr. 23.
S. Otter (1995), *Assessing Competence: the experience of the EHE Initiative,* in A. Edwards and P. Knight, *Assessing Competence in Higher Education,* London: Kogan Page.
W. G. Perry (1970), *Forms of Intellectual and Ethical Development in the College Years*, New York: Holt, Rinehart and Winston.
J. A. Petch (1953), *Fifty Years of Examining: The Joint Matriculation Board 1903 – 1953*, London: Harrap, pp 152 – 56.
C. Pollitt (1987), The Politics of Performance Assessment: lessons for higher education, *Studies in Higher Educatio*n 12, 87 – 98.
A. Powell and B. Butterworth (ca 1972), *Marked for life*, published privately.
D. E. Powers and M. E. Fowles (1999), Test-takers' judgments of essay prompts: perception and performance, *Educational Assessment* 6 (1), 3 – 22.

P. Race (2001), *A Briefing on Self, Peer & Group Assessment*, York: LTSN Generic Centre.
C. Rivlin and C. Roberts (2001), *Degree classification: a matter of equity*, Annual Conference of the Society for Research into Higher Education, 2001.
D. Rowntree(1977), *Assessing students: How shall we know them?*, (first edition) London: Harper and Row.
R. Simpson (1983), *How the PhD came to Britain*, Society for Research into Higher Education.
B. R. Snyder (1971), *The Hidden Curriculum*, New York: Knopf.
S Toohey (1999) *Designing courses for higher education*, Society for Research into Higher Education and Open University Press.
P Trowler and K Hinett (1994), Implementing the recording of achievement in higher education, *Capability* 1(1), pp. 53 – 61.
D Warren Piper(1985), Enquiry into the role of External Examiners, *Studies in Higher Education* 10, 331 – 342.
D. Warren Piper (1988), Staff development in universities: should there be a staff college?, *Higher Education Quarterly* 42, 238 – 252.
D. Warren Piper (1994), *Are professors professional? The organisation of university examinations*, London: Jessica Kingsley.
W. F. Williams (1979), The role of the external examiner in first degrees, *Studies in Higher Education* 4, 161 – 168.
R Winter (1993), Education or Grading? Arguments for a Non-subdivided Honours Degree, *Studies in Higher Education* 18, 363 – 378.
R Winter (ed) (1994), *Conference on the Future of the Classified Honours Degree,* Anglia Polytechnic University and Society for Research into Higher Education.
R Wood (1991), *Assessment and Testing,* Cambridge University Press.
H Woolf and D Turner (1997), Honours classifications: the need for transparency, *New Academic* 6 (3), 10 – 12.
M Yorke (1998), The Management of Assessment in Higher Education, *Assessment and Evaluation in Higher Education* 23, 101 –116.
M Yorke (2001), *Does grading method influence honours degree classification?*, British Educational Research Association Conference, Leeds, September.

General references

M. Atkins, J. Beattie and W. B. Dockrell (1993), *Assessment Issues in Higher Education*, Employment Department
J. Biggs (1999), *Teaching for Quality Learning at University*, Society for Research into Higher Education and Open University Press.
P. M. Broadfoot (1996), *Education, Assessment and Society*, Open University.
G. Brown, J. Bull and M. Pendlebury (1997), *Assessing student learning in Higher Education*, London: Routledge.
S. Brown and A. Glasner (eds) (1999), *Assessment Matters in Higher Education*, Society for Research into Higher Education and Open University Press.
J. Heywood (2000), 'Assessment in Higher Education', London: Jessica Kingsley.

Appendix: Good Assessment Practice

[Adapted from part of the University College London Good Practice Guide on Assessment, 1999, and based on L. Elton (1982), 'Assessment for Learning', in Bligh, D. (ed), 'Professionalism and Flexibility in Learning, Programme of Study into the Future of Higher Education' (The Leverhulme Study), Society for Research into Higher Education, Vol 6, pp. 106 – 135.]

Content

A. Aims, objectives and methods

(a) Purposes of Assessment
(b) Programme objectives
(c) Different methods of assessment
(d) Student motivation

B. Assessment as a measuring instrument

(a) Standards of measurement
(b) Reliability
(c) Validity
(d) The relationship of reliability and validity
(e) The 'backwash' effect
(f) Connoisseurship
(g) The use of weighting
(h) To grade or not to grade

C. Constructing examination papers.

(a) Structuring an examination paper
(b) When to set questions
(c) Essay questions
(d) Progressive questions
(e) Multiple choice questions
(f) Equal opportunities issues

D. Recent developments

Good Assessment Practice

Reasonably firm answers are given to some of the questions raised, but this is by no means so in all instances. It is therefore necessary to read what follows with a constructively critical mind. It is also necessary to appreciate that there may be a real tension between some of the matters advocated and the practicalities of assessment under particular circumstances, as carried out by busy academic staff.

A. Aims, objectives and methods

(a) Purposes of Assessment

Assessment can have a number of different purposes:

1. Selection and/or grading
2. Maintaining standards
3. Motivation of students
4. Feedback to students
5. Feedback to teachers
6. Preparation for life
7. Licence to practice.

There may well be others and not all of them can be achieved with a particular assessment method. Different purposes may need different methods

We believe that the overall aim of any educational programme should be that it leads to student learning. This may be obvious, but what is less obvious is whether assessment has a part to play in this. Purposes 3 and 4 help to promote learning directly, purpose 5 does so indirectly, purpose 6 does so provided one of the aims of the programme is preparation for life, but purposes 1, 2 and 7 do not in themselves encourage learning at all.

The purposes of assessment thus polarise into *assessment for learning*, also known as *formative assessment*, and *assessment for decision making*, also known as *summative assessment*. It is possible for an assessment to have both formative and summative aspects, but when the summative aspects are dominant, as they are for instance in most examination papers, formative aspects frequently get ignored. On the other hand, much course work assessment can be both formative and summative.

(b) Programme objectives

It may be obvious that what should be assessed in an educational programme is the extent to which its objectives are achieved, but in practice this is often not the case. Possible reasons for discrepancies between objectives and assessment are:
1. The programme has both process and product objectives, but only the product objectives are assessed; eg in laboratory work, the development of skills is a process objective and the production of a report a product objective, but only the latter is assessed.
2. The programme has both knowledge and skills objectives, but only knowledge objectives are assessed; eg a term essay on a particular topic also develops essay writing skills, but only the topic is assessed.
3. Some objectives are more difficult to assess than others, and only those more easily assessable are assessed (see validity and reliability below)
4. No explicit attempt has been made to link assessment to programme objectives; eg it is not possible to tell from an examination paper which objectives are being assessed.

(c) Different methods of assessment

Different assessment methods are appropriate for different assessment purposes and for the assessment of different programme objectives. Here is a list of possible assessment methods (others may be added):

- unseen examinations, structured or unstructured, with or without choice of questions
- open book examinations
- examinations with advance information about questions
- single question unseen paper
- multiple choice/ objective tests [a good guide is G Isaacs (1994), Multiple Choice Testing, HERDSA Green Guide 16]
- coursework assessment
- oral examination
- assessed report, dissertation, thesis
- assessments, the forms of which are negotiated in advance between examiners and students
- self and peer assessment
- group assessment

Some of these methods overlap, eg negotiated assessment could be in connection with an unseen examination, in which students are allowed to have a say in what is being examined..

Different methods suit different candidates better or worse, so that a mix is fairer than a single method. The introduction of course work assessment has certainly given coursework a deservedly higher profile, but it has to be accepted that its assessment is likely to be less reliable (for a discussion of reliability see below), since the work is not carried out under examination conditions. Since examination conditions are highly artificial, this feature of coursework may actually be considered an advantage, where 'natural' conditions are important.

Coursework assessment is usually thought to have both formative and summative properties, and this is true, but not unproblematic. If coursework assessment is preceded by wholly formative practice assessments, then students often fail to take advantage of the formative ones; if it is not, then students lack practice when they take the 'for real' assessment. One way of getting over this problem is to make every course work assessment count, but give two deadlines. Students can choose whether to hand in work before the first deadline, in which case it is commented on and returned for improvement, or not. Everyone's

work must be handed in before the second deadline, which is the assessment deadline, whether the work had been handed in before or not.

Quite generally, examiners might consider whether to give more choice to students, since choice - even between the frying pan and the fire - tends to motivate. Here is an example, which resulted from the first student riots are Berkeley in California. These were largely started through a demand to reduce the pressure of the 'grade point average' and were defused by giving students choices between being graded or simply passed. The new arrangement was that students were assessed on, say, six pieces of work or six papers, but that only three were graded,while the other three had simply tobe passed. This gave students the opportunity to shine at what they thought they were better at and into which they had put more effort.

Coursework assessment is particularly suitable for assessing process objectives, and essential for assessing true creativity and genuine problem solving abilities, neither of which can normally be assessed under the stresses and time pressures inevitable in formal examinations.

Oral examinations appear to be used for three very different purposes:

to test verbal skills and the ability to argue a point
to verify that students' written work in, say a dissertation, is their own
to help to make decisions on border line cases.

The first two purposes are legitimate, the third is more doubtful, since a brief and stressful oral, often conducted by an external examiner whom the student has not met before, is unlikely to lead to reliable decision making. **What is essential for all oral examinations is that examiners must receive training in interview and interpersonal skills.**

Negotiated, self and peer assessment moves assessment from wholly teacher determined to a joint enterprise between teachers and students. It is thus particularly appropriate for student centred programmes, but both teachers and students have to develop appropriate skills and attitudes before such forms of assessment are used.

Group assessment presents two particular problems:

How to assess the work of the individuals in the group, if not all contributed equally to the product which is to be assessed

How to assess the group process.

The first problem can be tackled in several ways:

Students are given equal marks, irrespective of their contributions
The individual student contributions are elicited through individual vivas (not really recommended)
Students are given a total mark and asked to divide it appropriately between themselves (in this and the following method it is highly desirable to negotiate appropriate criteria in advance of the work which is to be assessed)
Self and peer assessment are used at least in part.

The assessment of the group process could in principle be done through teacher observation, but this tends to disrupt the process. A much better way is to rely on the peer assessment of the participants, in which case it is essential to negotiate appropriate criteria in advance of the work which is to be assessed.

(d) Student motivation

Thirty years ago the commonly held view was that student should be motivated by their love for a subject and that examinations were a necessary evil, which detracted from this desired motivation. It is likely that few students ever held this view and that few teachers held it when they were students. More recently, it has been generally accepted that - in the words of an American author - 'grades are the campus currency', ie just as people at work are motivated by earning money, so students are motivated by earning marks. However, while people in work would not work without being paid, their motivation is often the interest of the work. Similarly, **it has been found that once students are satisfied that they are being fairly prepared for their examinations, their motivation is governed more often than not by the intrinsic interest of**

their study. Thus money and marks act as a trigger - without them there is no intrinsic motivation, but once the trigger operates, motivation depends much more on intrinsic interest than on size of the reward, whether in money or marks.

When the objectives being assessed diverge from the declared learning objectives - and this is distressingly common - students find themselves in a classical double bind situation. They either please their teachers through their intrinsic interest and fail their examinations or they concentrate on passing the examinations and displease their teachers. Sensible students choose the latter course, those that choose the former may well eventually take their exams from their hospital beds.

B. Assessment as a measuring instrument

Any measuring instrument involves a comparison with a *standard,* and aims to be *reliable* and *valid,* concepts which will be explained below in connection with assessment. In addition, any measurement is subject to the *Heisenberg Uncertainty Principle,* according to which the act of measurement changes what is measured in a not wholly predictable manner.

(a) Standards of measurement

In scientific measurement, a measurement is a comparison between the result of the measurement and an absolute standard, eg one compares one's weight as read by a weighing machine with the standard kilogramme in the international Bureau of Weights of Measures in Paris. In educational measurement, we assess students either against some pre-specified performance criteria or against the performance of a representative group of comparable students. The former, which is called *criterion referenced*, suffers from the same problem as the specification of objectives; ie it can suffer from both over and under specification. It is rarely used in university examining, except in certain professional examination, such as pharmacy. The latter, which is called *norm referenced,* depends on a sufficiently large and representative group of students being available. This may be the case at A-level, although even there the norms are not the sole criterion, but is effectively never the case in universities. The normal practice in universities would appear to be a mixture of criterion and norm reference. The borders between classes are permanently fixed, eg a First is above 70%, which implies that this mark satisfies the criteria for a First. This conclusion is however modified by current groups of students being compared for their quality with those of previous years, and the proportion in each degree class then being adjusted accordingly. This procedure, which can hardly be described as precise, relies considerably on standards from year to year (it may be noted that 'standard' is used here in a very different sense from that used earlier in this section), maintained through the memories of the examiners. While this does result in a similar degree class distribution for a particular degree examination in a particular university from year to year, it can and does lead to very different degree class distributions for a particular degree examination in different but comparable universities.

The belief that it is possible to fix the same class boundaries numerically also for different subjects is, however, wholly invalid, if marking practices differ substantially between different disciplines. It has been known for a very long time that mathematical type marking usually stretches over the whole scale 0 - 100%, while essay type marking more typically ranges from 30 to 70%. The argument for this difference, that it is possibly to reach perfection or total failure in mathematics but not in essays is absurd; both in mathematical and essay type assessment, perfection and total failure ought to be defined in terms of what is reasonable for the examination in question. Present indefensible marking practice is responsible for both the lack of Firsts in subjects like sociology and history, and the higher proportion of failures in subjects like physics and engineering. It also leads to absurdities in modular and multi subject degrees where it is essential for class distributions to be comparable in different subjects. Fortunately, most modular degree programmes now have marking schemes specified in terms of comparable achievements, so that the distribution of marks for different subjects is similar.

Finally, it is important to remember that class boundaries were settled a long time ago when virtually all assessment was based on finals papers. To take them over into schemes where a substantial part of the assessment is by other means is quite indefensible (rather like keeping the marks on a thermometer the same, but change from mercury to alcohol). In particular, it is well known that on average course work assessment leads to higher average marks and smaller spreads of marks. It is very likely that the apparent grade inflation over the past twenty years is due largely to examiners' ignorance of simple principles of measurement and is not a reflection of either improved learning or greater lenience in marking.

(b) Reliability

There are two kinds of reliability, which a measuring instrument must satisfy:

1. Two people who use the same instrument to measure the same thing should get the same result (examiner reliability)
2. Two supposedly equivalent instruments should give the same result when measuring the same thing (test reliability)

Innumerable investigations have shown that educational assessment at best is only moderately reliable in the first sense. Test reliability is far more difficult to verify, since it requires two supposedly equivalent tests to be given under the same conditions to the same students. However, it has been investigated, with the result that by and large the best students came out best on both tests and the worst students worst on both, but the ranking of middling students was very different in the two tests investigated. This is particularly important for multiple choice tests, which are 100% reliable in the first sense, but can be very unreliable in the second.

(c) Validity

An assessment is valid, if it assesses what it is intended to assess, which is usually specified in terms of the learning objectives which are to be assessed. Since the performance which is assessed through any form of assessment is inevitably different from the corresponding performance under more normal circumstances, the validity of the assessment can only be gauged by appropriate experts and can never be 100%. Points which experts will look for are that the assessment must fairly reflect the programme objectives which are to be assessed, ie it must not be testing

- outside the programme objectives
- too selectively within the programme objectives
- at an inappropriate level of the programme objectives.

One way to apparently ensure greater validity is to specify objectives very precisely, in the extreme in behavioural terms. However, this distorts the learning that is being assessed, since it focusses it in a most constraining way. Good learning is always more than being able to jump through pre-specified hoops, however well the hoops are pre-specified. **It is generally agreed that learning objectives must not be specified either too tightly or too loosely, if the learning is to be validly assessed.**

The validity discussed so far is called *'face validity',* because it involves the assessment being faced by an expert. There are other kinds of validity, which are concerned with the purposes of assessment, such as *predictive validity* in connection with future performance and *teacher feedback validity* in connection with improved teacher performance. In contrast to face validity, these other forms of validity are testable.

(d) The relationship of reliability and validity

Unfortunately, reliability and validity are not mutually independent from each other. Programme objectives which can be tested most reliably have two outstanding features:

- they largely test memory, since this leads to greater agreement between examiners than the assessment of higher learning objectives, eg any kind of understanding
- they treat all candidates equally.

Unfortunately, these are the very objectives which are of comparatively little importance in degree programmes, where it is usual to expect higher abilities and skills to be developed, and where increasingly learning programmes differ for different students. **Thus there has to be a trade off between reliability and validity.** The alternative, ie to test for memory and to treat everyone the same in the assessment, even though that is not what the programme objectives demand, is not however unknown. Its effect will be discussed in the next section.

(e) The 'backwash' effect

It has already been indicated that the Heisenberg Uncertainty Principle, according to which measurement changes what is measured is of great importance in assessment. It is even more important here than in physics, because when a measurement affects people, it affects their behaviour also before the measurement takes place. This is the 'backwash' effect, according to which **students' learning is guided by the assessment to come and the objectives being assessed become the students' learning**

objectives. In particular, if assessment assesses memory learning, then students tend to engage in memory learning, whatever their teachers may say. (This point was made already under 'student motivation'.)

The consequences of the backwash effect lend strength to the argument that high validity is more important than high reliability, although this is no excuse, once high validity is assured, for not doing one's utmost to increase also reliability. However, the result is almost inevitably a less 'fair' assessment, because fairness appears to demand that everyone is treated the same and is assessed in a highly verifiable manner. Perhaps the biggest difference between the educational and the so-called real world is this insistence on fairness in the former as a criterion above all others in assessment. Once it is appreciated that life is not fair and that education is a preparation for life, it will perhaps become more acceptable to take fairness off its high pedestal in the interest of better and more relevant education.

(f) Connoisseurship

The unreliability of assessment is particularly pronounced in disciplines with a strong creative and/or aesthetic component, but if it is accepted that all good assessment is to some degree unreliable, as regards both standards and individual performance, then perhaps it may become acceptable to use the concept of connoisseurship also in more usual disciplines. Connoisseurs are persons who, through training and experience, can make expert and reliable judgments in their specialist fields, which is not a bad definition of a good external examiner. Their judgment is accepted, because it is seen as both expert and reliable, and should therefore provide appropriate assessment standards. However, external examiners should normally have explicit checklists, so that their judgments are not totally holistic. A better model for the examiner may therefore be the judge in a sporting competition, such as ice skating or gymnastics, where both examiners and examinees know the dimensions on which the judgments will be made.

(g) The use of weighting

It is not uncommon to find that forms of assessment that are inherently of low reliability are given a small weight in the overall assessment, so as to reduce the effect of the unreliability. This is thoroughly bad practice, especially as very often these less reliable assessments are necessarily used to assess the more important learning objectives. Weighting should reflect the importance given to learning objectives, not the reliability with which they can be assessed.

(h) To grade or not to grade

Increasingly, the grading of degree work along a single dimension is being called into question. Is it really meaningful to describe three or four years' work in terms of a single number? If not, then it is time that the reporting of degree results were done in terms of a profile, in which some assessments would be on the traditional classified basis, some would be pass fail, some would be brief reports in words and some might be no more than a certification of attendance. Only such variety could meet the needs of different learning objectives in the years to come. It is important to realise that the American 'transcript' would be little more than a cosmetic improvement. Admittedly, it would give more information than a single class does, but it would still be in terms of the same kind of classification, whatever the learning that is being assessed. And as it is all too easy to calculate from it the 'grade point average', we might well be back with the single dimension in the end. Incidentally, reporting in terms of a profile also gets over the problem of different weights given to different assessments.

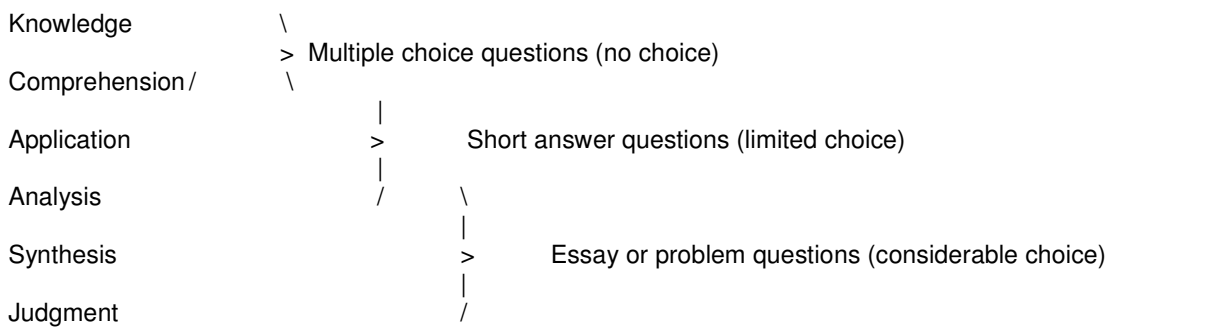
C. Constructing examination papers.

(a) Structuring an examination paper

The practice of unstructured examination papers, in which students are asked to select a portion of the questions set, which are ostensibly of the same difficulty, is only sound if the knowledge being tested is of very minor importance and if the academic skills tested by the different questions are of comparable difficulty. A broadly speaking hierarchical scale of such skills (developed from the work of B S Bloom) might be as follows:

- Basic knowledge, ie little more than memory work
- Comprehension, eg simply rearranged knowledge
- Application of knowledge, eg to practical problems
- Analysis, eg relation to basic principles
- Synthesis, eg bringing together relevant disparate considerations
- Evaluation, ie judgmental treatment of knowledge, analysis and synthesis.

Clearly, **basic knowledge should be treated comprehensively and not selectively, while the higher skills, which are concerned with academic skills applicable to the basic knowledge, can be exhibited with any basic knowledge. It is therefore reasonable that students should have a choice of knowledge base in which they demonstrate their academic skills.** This leads to the idea of a structured examination paper, with progressively more choice (or of course to different types of papers for the different levels of he hierarchy). A structured paper might look as follows:



(b) When to set questions

Many academics set questions only after they have completed a course for their students, partly so that the choice of questions will not influence their teaching and partly because they may not be certain until the end as to what they may not have time to teach and therefore should not be in the examination. The first consideration is laudable, but misguided, there is likely still to be a close relationship between teaching and questions set. The second consideration works on the not to be encouraged assumption that only that can be expected to be learned which has been explicitly taught, usually in lectures, rather than also that acquired from eg guided reading.

Others set questions well before they start their course, so as to distance them from their teaching. Unfortunately, this is likely to lead to stereotype questions, ie the very obvious ones which one is likely to think about when one is not close to the teaching.

The third way is to set questions as one proceeds with a course. This procedure is most likely to lead to the best questions and, provided one is aware of the dangers in teaching to the questions and avoids them, almost certainly the best.

Whichever method is used, it is desirable to set rather more questions than are needed and then to construct an examination paper which balances the various requirements, in particular the balance between the different items in the skills hierarchy.

(c) Essay questions

Particular care must be taken with essay questions, which usually contain an operational word, such as ‘discuss’, ‘compare’, etc.[For a more thorough treatment, see L Hamp-Lyons and B Heasley (1987), ‘Study Writing’, Cambridge University Press, pp. 140 – 142.] Such words should have definite meanings, which are shared through previous preparation by teacher/examiners and students. Unfortunately this is rarely the case, the understanding being generally tacit for the teacher or examiner and for that reason not shared with the students. There is no way of knowing whether this tacit understanding is the same for all examiners, and so only those students who happen to share the tacit understanding of a particular examiner are likely to do well. What is needed for a fair assessment (here fairness is both possible and desirable) is for that understanding to become explicit between teachers and students and for examiners to agree to it. Possible operational words are:

- define
- describe (also: list)
- analyse
- explain (also: account for)
- assess
- compare
- contrast
- argue (also: justify)

- critically examineevaluate

and some of these are explained more fully in the attached article. They have different meanings and they correspond to different levels in the hierarchy of academic skills. All this must be taken into consideration, when constructing a balanced examination paper.

The perhaps most common operational word, ie 'discuss', is not in the list, as it has too many different meanings in different circumstances, each one of which is better expressed by one of the words in the above list. Could it be that the popularity of this word in examination questions reflects the not uncommon attitude in examiners to deliberately be vague in a question, in order to see how students tackle it and then match their responses to the examiner's preconceived perceptions, which are of course unknown to the examinees? [Note that 'Study Writing' does allow 'discuss', but calls it 'one of the most difficult types of essay question'.]

There is also the essay question, which is literally a question. Here are three such questions, which all were alternative questions in a genuine paper:

> Who were the "New Karaites"?
>
> How has the excommunication of Spinoza been explained?
>
> How significant was the Thirty Years War for the Jews in Western and Central Europe?

Without any familiarity with the subject, it would appear that the first calls for a purely factual answer, while the second expects an evaluation of different possible explanations, and the third could legitimately be answered with the single word: "very". This is an indication that **questions which are literally questions are in general unacceptable.**

Finally, it is worth raising the issue of using computers in essay examinations. Students are now strongly encouraged to word process their course work. Is it then right that they should return to pen and paper for their examinations? And if not, can the problems associated with bringing computers into examinations be overcome? There are at present no answers to these questions, but they are worth asking.

(d) Progressive questions

Questions which are in several parts, and in which information in the earlier parts helps in responding to the later ones, are common in the sciences. In such composite questions, the first part is often a purely memory question, the second part uses the information contained in the first part in solving a problem and the third part may similarly use it for a more theoretical question. In such cases, the wording of the first part gives information on the later parts, eg what might be a good starting point for solving a particular problem, and thereby makes the second part easier than it would have been, if the first had been omitted. Since a most important aspect of problem solving lies in the identification of a suitable starting point, this may devalue the question as a test of problem solving skills. Furthermore, if such a question is now set in an open book examination and the first part is omitted, as being purely memory work and therefore unsuitable for an open book question, then the problem part may have been made considerably harder by this omission, a fact that is often not appreciated.

(e) Multiple choice questions

The perception that MCQs can test only for knowledge recall is doubly wrong: all MCQs test for recognition rather than recall, but they can test at all levels of the hierarchy discussed in section C(a). However, the higher the level to be tested, the more difficult is it to set good questions. Quite generally, **just because MCQs give no information concerning students' thought processes, it is particularly important that they should be of a high professional standard.** Examiners should not use MCQs, unless they have been trained in their use, and this is true even if the MCQs are taken from professionally generated question banks.

(f) Equal opportunities issues

This may be a good point to refer to equal opportunities issues. The following points ought to be considered in connection with any questions set in an examination:

- Is there a race or gender bias in the question?

- Have the needs of disabled students been adequately catered for?
- Is the language chosen such as not to unduly handicap those for whom English is not their first language?
- Has some inappropriate cultural bias been introduced?
- Has the anonymity of scripts beeen assured?

D. Recent developments

The matters discussed so far, with the possible exception of self and peer assessment, all relate to well established practices. **Recent developments have concentrated on forms of assessment, which take account of the movement to use assessment in order to encourage active and student centred learning, which make it difficult to assess all students in the same manner, as well as on the necessity in these days to make assessment more cost effective.** An excellent and easily readable book, which brings traditional concerns up to date, is S Brown and P Knight, *'Assessing Learning in Higher Education'*, Kogan Page 1994. This should be read in conjunction with S Brown and B Smith, *'Getting to Grips with Assessment'*. SEDA Special No 3, Staff and Educational Development Association, 1997, which gives practical advice and check lists. Both are strongly recommended, as is the booklet *'Strategies for diversifying assessment in higher education'* by S Brown, C Rust and G Gibbs, The Oxford Centre for Staff Development, Oxford 1994.

Chapter Three

Some Basic Assessment Dilemmas with Particular Reference to Portfolios

Brenda Johnston

Section One - Introduction

A basic assessment dilemma in UK higher education today

The changing context of higher education has entailed questioning of traditional assessment systems in many quarters (although apparently not much action, as described in Chapter Two). The changing context has, from our point of view, two major aspects.

On the one hand, in modern day society, educational qualifications have an increasingly important role in creating and limiting employment chances (*Editorial* 1998, p.302). There are larger numbers of students than previously. The result is that there are pressures on higher education to produce summative assessment systems with high levels of consistency and which are easy and economical to implement (p.306). They must also be usable in terms of 'feedout' to employers and postgraduate schools (Knight 2001 p.15).

On the other hand, another range of pressures focus attention on the need for assessment that is formative and developmental in orientation. There is a major movement towards recognition of the nature of learning and attempts to align practice with theories of learning. The work of Marton and Saljo (1984) on deep and surface learning; of Kolb (1984) on experiential learning; of Mezirow (1981) and Knowles (1970) on andragogy; of Schön (1983, 1987) on reflective practice; and of Vygotsky (1962, 1978) on the dynamic social nature of learning have been particularly influential here. Moreover, it has been recognised that there are strong links between learning, feedback and assessment (see Black and Wiliam 1998; Madaus 1988 [school-level education]; Becker, Geer and Hughes 1968; Brown and Knight 1994, Elton and Laurillard 1979; Hinett 1997, and James 1997 cited in Broadfoot 1998; Snyder 1970 [for higher education]). In their review of how assessment can best support learning, Black and Wiliam highlight the importance of intrinsic motivation, confidence building, detailed and substantive feedback, collaboration rather than competition and the need to encourage students' metacognitive skills and so ability to monitor and direct their own learning. These pressures for formative and developmental assessment, in line with learning theory, have been coupled with demands for supposedly "new" and varied skills and abilities (such as negotiating and team skills), together with renewed interest in "old" skills (such as critical thinking) in higher education graduates and lifelong learners (e.g. CVCP 1998; DfEE 1997). An additional pressure is that the student population has become increasingly diverse and requires more varied approaches to teaching, learning, and assessment.

This second set of pressures is not readily consistent with the demands for reliable, convenient, economical high-stakes assessment. Conflicts between these different pressures are inevitable.

Plan for this chapter

This chapter seeks to explore various aspects of the basic dilemma described. It does this through a close investigation of one assessment method, portfolio assessment. However, many of the points made in connection with portfolios can be transferred to other assessment methods, including those often associated with traditional and innovative teaching and assessment. In order to achieve this close examination, we will review the research evidence critically, including its theoretical and methodological underpinnings. Indeed an examination of the theoretical and philosophical underpinnings of the current research debates is a central issue in this chapter. Such an examination is essential in order to throw open assessment practice and research to critical scrutiny. We hope that by exploring the theoretical debate through the concrete example in the shape of portfolio research that the discussion will be grounded in aspects and dilemmas that readers can recognise.

This report is not a comprehensive review of research in the field of portfolio assessment, given time limitations and the now vast and diverse nature of the literature. Instead, it aims to highlight significant findings as well as substantive and methodological gaps in the research.

The report focuses on two major areas:

- agreement over rating in portfolio summative assessment
- claims about learning achieved through formative assessment and reflection when constructing portfolios.[8]

It also deals briefly with the practical issues involved in implementing and running a portfolio assessment scheme. The first of these topics relates particularly, but not exclusively, to the first set of pressures described above, those concerned with achieving consistent outcomes. The second relates to the second set of pressures described above, those concerned with achieving learning.

In order to explore the first area, the report begins with a discussion of two fundamentally different approaches to assessment: traditional approaches arising out of psychometrics and; more recent approaches arising out of interpretivist assessment. Understanding of these two approaches with their different theoretical and philosophical bases is central to an understanding of a critique of the literature on agreement over rating of portfolios for summative assessment. The report then discusses empirical research on agreement over rating from the two different traditions (positivist, psychometrically-based and interpretivist). It reviews specific articles and draws out overall conclusions as well as gaps in the literature. After this, the report discusses the complex links between portfolio assessment and learning from a theoretical point of view. It then reviews specific articles, drawing out overall conclusions and pointing out gaps in the literature.

The report will consider research and theory from different countries and different sectors of education, rather than just higher education in the United Kingdom. Many of the principles underlying research outcomes elsewhere are relevant to the higher educational context here. Moreover, much of the significant research on portfolios has been done in areas other than British higher education. One particular point to note is that much thoughtful and useful research on portfolio summative assessment has been done in the field of writing in the United States. This is a huge field there with a thriving research community, as well as active teaching departments, in the areas of the teaching, learning and assessment of writing. Researchers have investigated writing from school through to postgraduate levels and beyond. This research will be mentioned extensively in the section on agreement in portfolio summative assessment.

One minor point to note is that in the US literature and in some of the quotations in this report, the terms “evaluation” and “assessment” are used interchangeably. US spelling conventions have been preserved when quoting from North American texts.

Although it is impossible to say how widespread portfolio assessment is, the research indicates that it has been adopted in many areas of the curriculum (e.g. medical education, school and university teacher training, writing, mathematics). In some areas such as architecture and art and design, portfolio assessment has been commonplace for many years. In university teacher training and accreditation, Baume and Yorke (2002) report that since 1992, the UK Staff and Educational Development Association (SEDA) has recognised 60 programmes using portfolios in various countries and that about 1700 teachers have successfully completed these programmes.

Much that is written about portfolio assessment, as for other assessment methods, consists of advocacy and guidelines for the introduction of portfolio assessment. Indeed, in 1994 Herman and Winters found that “of 89 entries in portfolio assessment topics found in the literature over the past 10 years, only seven articles either report technical data or employ accepted research methods. Instead most articles explain the rationale for portfolio assessment; present ideas and models for how portfolios should be constituted and used; or share details of how portfolios have been implemented …” (p.48). In 1996, Wade and Yarborough found only nine research studies of the use of portfolios in teacher training (p.64). The purpose of this report, however, is to focus on the findings of research which, thankfully, is now far more extensive than in 1994 and 1996.

Another point to note is that throughout this report we refer to frequently to “students” and “teachers”. We are using these terms loosely. We recognise that “teachers” may be “university teachers”, “school teachers”, “professional mentors” and so on. “Students” may be “school pupils”, “university students”, “doctors” or other “professionals”.

Why do people advocate portfolios?

Portfolios are seen as having the potential to:

[8] See Glossary *p.95* for explanation of terms such as formative assessment, summative assessment.

- engage students in tasks which are central to the educational process as perceived by theories of learning
- encourage students to take an active role in their own learning in the shape of formative assessment
- offer "authentic" assessment which in turn is likely to provide predictive information about how a student will perform after moving beyond the assessment
- allow assessment of a wide range of learning achievements, providing detailed evidence of these which can inform teaching as well as enabling formative assessment
- help students develop reflective capacity which will in turn enable them to continue learning after passing beyond the immediate course
- encourage students to take an active role in their own assessment in that they may be able to select which work goes in the portfolio
- track students' development over time
- showcase students' responses to a wide range of assignments (*Editorial* 1998, p.303; LeMahieu, Gitomer and Eresh 1995, p.11; Murphy 1994, p.179-80).

Various issues have to be unpacked in examining these claims. The reader will note that most claims for the benefits of portfolios relate to the claims made about the learning they can promote. Many of the problems raised about portfolio assessment relate to the consistency and fairness of their assessment.

What is a portfolio? What are the purposes of portfolios

At a basic level, a portfolio contains a collection of items for assessment rather than a single piece of work. However, beyond this portfolios take various forms and have various purposes and so contain different types of materials. Moreover, the design, use and contents of portfolios reflect different conceptions about the underlying aims of the curriculum (Murphy 1994). Frequently people, including researchers in the field, only seem to be aware of one type of portfolio, or at least discuss one type as if it were *the* only type. It is important to be aware of the range of portfolios and the pedagogic and assessment implications of different types.

Murphy, speaking within the context of the teaching of writing in the United States, suggests three types of portfolios: behaviourist, "structure of the disciplines" and cognitivist-constructivist.

- In a *behaviourist* portfolio the focus is on development of discrete sub-skills, their rather mechanistic practice probably through some kind of worksheet, feedback given until the skill is mastered, and systematic assessment of skill development (p.181-2, 185). We will not discuss this type of portfolio in this report. But as Murphy explains, it might consist of completed worksheets offering basic testing and practice of discrete sub-skills.
- In a *"structure of the disciplines"* portfolio the focus is on "subject matter, the way scholars and practitioners in the discipline understand its structure, its predominant modes of inquiry and its fundamental concepts" (Bruner 1971; Posner 1992 cited in Murphy p.185). Presumably, a portfolio which contained content oriented pieces of work, structured around particular modes of expression or activity in a discipline would fit this mould.
- In a *cognitivist/constructivist* portfolio, the focus would be on the learning of concepts and thinking processes. "Students engage in purposeful activities requiring critical thinking skills, problem solving, decision making and judgements" (p.185). Murphy argues that this last type of portfolio also connects with experiential learning concerns where the emphasis is on "the intellectual and social development of the individual" and learning is related to each student's own experience (p.185-6). This experiential focus is particularly clear in the teacher training portfolios discussed in *Section Three.*

Others might wish to suggest more categories and/or to dispute those above. Certainly some subject areas seem to have more affinity with certain types of portfolios than others. Presumably a portfolio might contain work representing a mixture of types. For example, a portfolio could contain content focused material, but this could have been produced in an approach which stressed the importance of learning through problem-solving and critical thinking.

Portfolios have different functions in different courses. The level of *individual student agency* in selecting work to go in the portfolios and design of their format varies widely according to portfolio type and the underlying assumptions of the curriculum. One critical, but somewhat fuzzy difference between portfolios is whether the portfolio is to be used as:

a) *a file (usually) of student work* (usually) produced during a course or some kind of work/professional experience. Portfolios in art and design and architecture courses are traditionally in this mode. The *individual* contents of the file may or may not have been offered up for formative and/or summative assessment before being placed in the portfolio. Each portfolio for members of a course may contain the same items to be assessed, as distinct from students selecting or producing different items. The items

may have been selected by the student in which case a Type a) portfolio is becoming closer to a Type b) portfolio. Important to note is that learning may well have taken place on the course, but not necessarily as a result of activities carried out in creation of the portfolio which has been a passive receptacle into which course products/drafts in various forms have been placed.

b) An *active learning tool* requiring analysis and review of the contents by student. In this approach, portfolios are, minimally, a tool to promote reflection. To this end, they may, as David Baume (2001) described, contain:
    - *evidence* (e.g. reports, essays, designs) as it appropriate in a particular discipline. This evidence may or may not have been produced for the portfolio. It may arise out of a project or doing another part of the programme etc.
    - *labelling, signposting, structuring of the evidence*
    - "*critical reflection or commentary*, very probably written especially for the portfolio .. which contextualises the evidence … makes sense of the evidence" (p.5-7).

    Baume and Yorke (2002) argue that:

    > … a portfolio typically includes evidence drawn from practice. Crucially, it usually also contains reflective commentary … in which the course participant show how he/she has interrogated his/her experience and related his/her practice and understandings to cognate evidence from the literature and elsewhere. It is typically expected that the portfolio will be scholarly, and that insights will go beyond a quotidian pragmatism to connect with the relevant theoretical constructs (in press).

    Some have suggested that portfolio construction should also require student participation in the assessment process (e.g. Paulson, Paulson and Meyer 1991; Rief 1990; Tierney et al. 1991). They argue that students should be involved in:
    - establishing guidelines for what the contents should be
    - selecting materials for inclusion in the portfolio
    - establishing criteria for the assessment.

    In such a case, as well as promoting reflection in the students, portfolios
    - "become personal collections of educational experiences over a period of time"
    - provide a very active means whereby students can participate in their own assessment
    - "provide a more equitable and sensitive portrait of what students know, and are able to do, than do traditional assessments" (Snadden and Thomas 1998 p.193).

In both cases a) and b), the portfolios may be offered up for formative and/or summative assessment (or potentially neither). The assessment may involve the person who produced the portfolio and/or peers and/or tutors and/or employers.

There is a wide range of practice concerning types and purposes of portfolio assessment. In some areas such as teacher training portfolios, student independence as well as responsibility and agency in selecting and designing content of the portfolio is essential. In other cases, where portfolios are merely repositories of work done, these aspects are far less important. Students have to complete a core curriculum and are judged on the standard of the work produced, although the way they tackle each individual piece of work may vary widely.

For further discussion of the different nature of portfolios, especially in relation to teacher training see Wade and Yarborough (1996, p.65). Allen (1995) discusses the wide range of portfolios forms and purposes in writing programmes in the United States. Allen, who co-ordinated a project across various universities in the United States where participants commented on portfolios from other universities and colleges, found that portfolios ranged widely in type, even within one subject area. Some were on paper while some were electronic versions. Some portfolios were used for placement purposes while others were exit portfolios. Some portfolios contained every draft while others only contained final papers. Some were part of large scale programmes while others were used within individual classroom only (p.72). Across all subject, frequently the portfolios discussed in the literature are exit portfolios, portfolios used for summative assessment at the end of a course. Challis (1999) discusses what form medical portfolios might take through description of existing examples (p.376-383).

We will now turn to our first major issue: agreement over rating of portfolios for summative assessment. This is an area of major controversy within the portfolio assessment literature.

Section Two Agreement over Outcomes in Portfolio Summative Assessment

Theoretical dilemmas and practical implications: positivist and interpretivist approaches to summative assessment

In order to follow the findings of the research literature on portfolio use in summative assessment, it will be useful to understand the philosophical and theoretical bases which underpin different approaches to assessment. These bases entail rather different practical approaches to assessment. This report explores two basic strands: (1) positivist approaches and (2) interpretivist approaches. The first of these has dominated assessment theory, although not necessarily practice which usually tends to be fuzzier than theory, until recently. The purpose of this discussion is not to argue that either approach offers all answers to all assessment problems, but rather to establish a framework in which existing research can be evaluated and to present two positions in order to clarify the choices that are available to assessors. We hope that this makes the assessment debate more informed and more sophisticated.

Positivist approaches to assessment

These approaches arise out of the positivist paradigm, with its belief in the possibility of 'objectivity', 'scientific' measurement and certainty. The assumption is that "student ability … is a fixed, consistent and acontextual human trait" (Huot 1996, p.550). A key notion is that there is an "ideal", objective grade the assessors should be seeking to discover. Key aspects in these approaches are the concepts of *reliability* and *validity.* As Brown, Bull and Pendlebury (1997) describes,

> The standard approaches to reliability and validity are derived from psychometrics, a subject which is particularly concerned with the development of personality and intelligence tests. The psychometric approach is based upon the notion of *an ideal which can be achieved if only one can reduce the errors* (p.233). [emphases mine]

This last point is central to discussions of reliability.

Reliability

According to *reliability* criteria, assessment should be:

- objective
- accurate (methods of measurement stable and sensitive)
- repeatable (measurement procedures must be clear and consistent from case to case)
- analytically sound – tests have to be correctly analysed and entered (Knight 2001, p.11)

As Moss (1994) explained:

> Typically, reliability is operationalised by examining consistency, quantitatively defined, among independent observations or sets of observations that are intended as interchangeable - consistency among independent evaluations or readings of a performance, consistency among performances in response to independent tasks and so on (p.6).

Feldt and Brennan (1989) describe the "essence" of reliability analysis as the "quantification of the consistency and inconsistency in examinee performance" (p.105 cited in Moss p.6).

Writers on assessment often make the point that reliability is only concerned with consistency and not with fairness in any other sense (e.g. Elton 1987; Huot 1996). Huot points out, for example, that we could assess writing by counting the number of words written and thereby achieve perfect reliability in scoring between raters. However, such a test would hardly be a "fair" assessment of the writing in the sense of it being a meaningful judgement about the quality of the writing.

A fundamental principle of positivist, psychometrically-based approaches is that "without reliability, there is no validity", although reliability will not ensure validity.

Validity

Validity is concerned with "fitness for purpose". There are different forms of validity, described and defined somewhat differently in different literature. These types of validity are overlapping, interconnected and frequently qualitatively different. I do not wish to explore these complications here so will merely set out some concepts used frequently in the literature. *Figure 1* below lists the various forms of validity commonly discussed in the literature.

Figure 1 Types of validity

Face validity	This is concerned with "the degree to which a test *appears* to measure what it purports to measure whereas the other forms of test validity … *provide evidence* that the test measures what it purports to measure. Thus, although face validity can never take the place of the other forms of test validity, it is still important because most people react more favourably to tests having high face validity" (Borg and Gall 1989, p.256).
Consequential validity	"This … is concerned with the consequences of the nature and load of assessment upon teaching and student learning and other aspects of a system such as administration and research" (Brown, Bull and Pendlebury 1997, p.239). As Miller and Legg (1993) describe: > The consequences of test use can be intentional or unintentional, as well as negative or positive. For example, in most alternative assessment, a potentially positive and intended effect is that the curriculum and instruction will benefit from a closer match to the test. Negative effects of testing have included teaching narrowly to the content of the test and wasting too much time on test preparation activities (p.9). Consequential validity has assumed greater importance in the wake of growing concerns about the "institutional and societal impact of a given assessment programme" (Broad 2000, p.251).
Intrinsic validity	"Do the assessment tasks measure the learning objectives of the course?" (Brown, Bull and Pendlebury 1997, p.241). The danger is that it may not be possible to express the objectives in measurable terms. Judgement rather than measurement is required. Over-specificity of objectives can lead to tutor fatigue in marking, and fragmentation of underlying aims of the course etc.
Construct validity	This is "the extent to which a particular test can be shown to measure a hypothetical construct" (Borg and Gall p.255) on which the test is based. For example, an assessment task might test "creativity" or "problem-solving". Constructs are not directly observable but rather are inferred on the basis of their observable effects on behaviour" (p.255).
Content validity	This is the degree to which the content of the assessment "represents the content that the test is designed to measure. … a test need not cover *all content* to be content valid, but must cover a *representative sample* of this content…. If test items cover topics not taught in the course, ignore certain important concepts, and unduly emphasise others as compared with their treatment in the course, the content validity will be lower" (Borg and Gall p.251).
Criterion validity	"Are the results of the assessment tasks comparable with those obtained on other assessments of known standard by similar groups of students?" (Brown, Bull and Pendlebury p.242). It is usually difficult to judge these matters in complex assessment tasks. For example, it is difficult to find a "similar" group of students. Concurrent and predictive validity are often considered to be different types of criterion validity.
Concurrent validity	"Does performance on the assessment tasks match that obtained by other assessments of the same group of students taken at roughly the same time?" (Brown, Bull and Pendlebury 1997, p.242). It is usually very difficult to judge these matters because of matters such as different ranges of marks in different types of assessment, different policies towards marking them and different levels and types of student performance in the different types of assessment (Brown, Bull and Pendlebury p.242).
Predictive validity	"Do the assessment tasks predict future performance accurately?" This is useful in work situations for example where someone wants to know whether performance on an assessment can predict how well a person can do a particular job. It is hard to find tasks and to do the empirical research that allow testing of this kind of predictive validity. For example, some employment related tasks may take years to work through and graduate performance will be affected by many factors other than original learning in higher education.

Complicating factors when judging validity, other than the formidable, inherent difficulties in measuring each type of validity, are that:

- just because an assessment method is valid according to one of the types above, it doesn't mean that it is valid for the others.

- an assessment may have *intrinsic* validity in the sense that the assessment method addresses the learning objectives of the course, but the *extrinsic* validity (in terms of the desirability of the objectives of the course) may be weak.

Important to note is that the nature and extent of validity of a test will have a "backwash effect" on what students prioritise in their learning and also in how teachers teach (See Becker, Geer and Hughes 1968, Broadfoot 1998; Brown and Knight 1994, Hinett 1997, and James 1997 cited in Broadfoot 1998; *Good Practice Guide* in the Appendix; Snyder 1970).

Conflict between reliability and validity in complex tasks

Reliability and *validity* are often in conflict (See e.g. Elton 1987, *Good Practice Guide* in Appendix). As Knight (2001) discusses:

> The need for reliability pushes us towards certainty and simplicity but modern higher education curricula value complex, fuzzy achievement exemplified by soft skills, autonomy, creativity, incremental self-theories, interpersonal fluency, etc (p.13).

He points out the dangers of this conflict:

> 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 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*" (p.13).

Broad (2000) describes the conflict between reliability and validity in assessment of writing. Although he is discussing assessment of writing in a composition programme in the United States, his comments are equally applicable to assessment of the various kinds of writing that happen in British higher education:

> This stark and unsettling controversy goes to the heart of teaching and assessing writing. Is the push for agreement a safeguard against unfairness and chaos? Or is it an oppressive effort to make readers read in peculiar and diminished ways? Does rejection of the goal of consensus among raters herald a new age in writing assessment in which evaluative disagreements become meaningless and educational? Or does it merely make an impossibility out of a difficulty at the expense of fairness to students and clarity of professional standards (p.216).

In more "valid" complex tasks, there are some difficulties with reliability as these require judgement in assessment. It is hard to achieve either inter- or intra- rater consistency. Some have recognised the necessity of following a "good enough" approach, rather than reaching exact and consistent agreements (e.g. Davis et al. 2001; Rossi and Freeman 1993b cited in Davis et al).

One view expressed by Brown, Bull and Pendlebury (1997), and probably widely accepted by many, is that although it is problematic to assume that there is one ideal for higher order knowledge and processes, the concepts of reliability and validity are nonetheless helpful for clarifying the purposes of assessment. They point out that educational assessment requires somewhat different processes from those required in psychometrics.

> … non-statistical approaches such as the use of judgement in the identification of appropriate tasks and content, the use of blueprints of learning objectives (outcomes) and assessment tasks and the translation of task performance into marks are required, as well as the approaches provided by statistical analysis (Brown, Bull and Pendlebury p.233-4).

Moves to expand traditional notions of validity

Concerns about authenticity in assessment and the social and educational consequences of assessment have resulted in recent years in the expansion of definitions of validity. Until then people did not pay much attention to validity. Newer notions of validity stress that a 'valid' procedure for assessment must have a positive impact on and consequences for the teaching and learning. A 'valid' test must take account of learning theory as well as empirical data in its design (Cronbach 1988 and Messick 1989 cited in Huot 1996, p.550-551). Recent validity researchers:

> … have stressed the importance of *balancing* concerns about reliability, replicability, or generalizability with additional criteria such as "authenticity" (Newmann 1990), "directness" (Frederiksen & Collins 1989), or "cognitive complexity" (Linn, Baker, & Dunbar 1991). This balancing of often competing concerns has resulted in the sanctioning of lower levels of reliability, as long as "acceptable levels are achieved for particular purposes of assessment " (Linn et al., 1991, p.11 see Messick 1992, and Moss 1992 all cited in Moss 1994, p.6).

However, typically these newer approaches to validity still require significant degrees of standardisation and reliability in a quantitatively consistent sense.

For the interested reader, introductory discussions of validity and reliability can be found in Heywood (2000) and in Brown, Bull and Pendlebury (1997). Earlier discussions include (Cronbach 1971; Hartog and Rhodes 1935, 1936; Milton, Pollio and Eison 1986 all cited in Heywood 2000; Elton 1987). Further discussion of reliability principles can be found in texts such as Crocker and Algina (1986); Cronbach 1990; and AERA et al. 1986. Reliability is also discussed in the *Good Practice Guide* in the appendix to Chapter Two.

Post-positivist approaches to summative assessment

Some, dissatisfied with the positivist principles underlying much assessment and associated problems, such as the basic conflict described above between validity and reliability, have advocated an essentially different approach to assessment, rather than tinkering at the edges of psychometrically-based assessment. They have argued for an approach to summative assessment based on post-positivist, interpretivist principles.

Such arguments have been strengthened because, as previously described, educators are increasingly searching for ways to encourage learning processes in line with learning theory. However, they are likely to be frustrated if the assessment of their courses promotes different skills and values from those encouraged in learning theory. It has been known for many years that high-stakes, summative assessment has a strong impact on teaching and learning practices. Early work by Snyder (1970) and Becker, Geer and Hughes (1968) documented this process in great detail at MIT and the University of Kansas respectively. (At MIT, it resulted in good students dropping out, because they did not want to play the exam game.) There is a large body of more recent evidence which have similar findings (e.g. Corbett and Wilson 1991, Hinett 1997, and James 1997 cited in Broadfoot 1998; Johnston, Weiss and Afflerbach 1990; Smith 1991 both cited in Moss 1994). This has caused concern in that standardised assessment does not foster intellectual activities such as creativity; individuality of approach; encouraging students to find their own purposes for learning; and encouraging teachers and students to work together to develop criteria and standards to develop their own work. These are all aspects typically encouraged in recent learning theories. Given the questionable value of much high stakes, summative assessment, Wolf et al. (1991) have argued, educators need to

> … revise our notions of high agreement reliability as a cardinal symptom of a useful and viable approach to scoring student performance [and] seek other sorts of evidence that responsible judgement is unfolding (p.65 cited in Moss 1994, p.6).

These moves relate to the pressures for "authentic assessment". Authentic assessment is the movement for assessments that test a student's ability to carry out tasks (e.g. problem solving, teaching) that resemble authentic situations. Portfolio assessment is one manifestation this. As Broad (2000) elaborates:

> The thesis driving authentic assessment is that complex social, intellectual, and rhetorical abilities cannot be validly measured through the radically simplified instruments of standardised testing used in the peculiar context of the examination room (Wiggins 1993). In fact, complexity and sensitivity to students' chosen contexts are considered crucial qualities in the process. While advocates for authentic assessment deprivilege such psychometrical requirements as interrater agreement, they claim to attend better to educators more important responsibilities (p.250).

Please note that much innovative thinking in the field of post-positivist assessment has happened in the assessment of writing. Many of the references below will be to writing. This does not mean that I am assuming that all portfolios will contain written work. Clearly art and design portfolios, to mention but one example, will not. I am assuming that comments made and about reading and writing and interpretation below are transferable to other type of tasks besides written ones. Interpretation can also apply to any assessment artefact or process which requires judgement of students' skills, understanding, creativity, originality, and problem-solving capacities (e.g. artwork in art and design; various kinds of project work in science and engineering).

The bases of post-positivist approaches to assessment

So what are the bases of post-positivist approaches to assessment? In order to explore this point, we will draw on two bodies of literature: evaluation (of institutions, individuals other than students etc); and hermeneutic, interpretivist assessment literature. We will use the terms – constructivist, hermeneutic and interpretive – which arise out of the same type of post-positivist approach to assessment. They each have their own particular nuances and origins, but a central feature is that realities, especially social realities, are seen as mental constructions or interpretations, rather than absolute, objective truths. "'Truth' is a matter of consensus among informed and sophisticated constructors, not of correspondence with an objective reality" (Guba and Lincoln 1989 p.44), an idea which, incidentally, goes at least as far back as the bible (St John Chapter 18, v38).

In this section, we draw the reader's attention to the fact that we are citing extensively from the US assessment literature which uses the terms "evaluate" and "assess" interchangeably. We are also, perhaps somewhat confusingly, drawing heavily on evaluation literature where evaluation refers mainly to evaluation of institutions, individuals rather than students etc. In a sense this does not matter, since much of what is applicable to other kinds of assessment/evaluation is also applicable or at least can be related to assessment of students. We hope that the intermingling is not confusing and will endeavour to clarify as much as possible.

Evaluation [of institutions etc.] practitioners, researchers and theorists (e.g. Simons 1987; Stake 1972; Stake and Kerr 1994) have long discussed and used post-positivist approaches to evaluation and many parallels can be drawn with assessment. Simons (1996), for example, draws attention to the complexity and need for interpretation involved in making evaluations. She argues that evaluations are constructed interpretations, rather than absolute facts. She cites Eisner (1993) who argues that "perception is a cognitive event and […] construal, not discovery, is critical" (in Simons 1996).

However, it is on the work of Guba and Lincoln, *Fourth Generation Evaluation* (1989) that the present analysis will concentrate. As Allen (1995) summarises:

> In Guba and Lincoln's approach, evaluation is "socially constructed" in a dialogue between evaluator and the person being evaluated, with a chance for all "stakeholders" in the evaluation to express and defend their viewpoints. Both the Moss [1994] and Guba and Lincoln schemes stress the importance of local context, connection and holistic integration over the distance, independent observation, and aggregated scores from separate evaluations which are utilized in psychometric assessment (p.68).

Guba and Lincoln criticise previous generations of evaluation research for their failure to acknowledge the plurality of values in making judgements and over-commitment to scientific principles. As they point out, almost every society:

> …turns out to be value-pluralistic. Then the question of whose values are to be taken into account, and how different value positions might be accommodated becomes paramount (p.8).

The role of context

In the types of evaluation advocated by Guba and Lincoln, *context* is central. "… absence of context distorts the ability of individuals who rely on it to make meaning" (Huot 1996, p.557). They suggest that the constructions people put on the situations they are in, the evaluations they make:

> … are inextricably linked to particular physical, psychological, social and cultural contexts within which they are formed and to which they refer (p.8).

At a practical level, assessors make judgements which are not absolute, but instead depend on aspects such as: the particular features valued by that assessor and probably a larger group according to their disciplinary and cultural backgrounds; the standard the assessor knows to expect from the assessed group; and so on.

Assessment is "a sociopolitical process".

> Social, cultural, and political factors should not be viewed as unattractive nuisances … that threaten validity, but as integral and meaningful components of the process, without which the evaluation effort would be sterile, useless and meaningless (p.253).

This tallies with post-positivist conceptions of "truth" such as those expressed by Foucault. Academia is supposedly concerned with the pursuit of "truth", but "truth" in Foucault's view is a socio-historical construction, mediated by particular discourse practices (Goodwin and Duranti 1992, p.31). As McCormick (1990) elaborates:

> Thus, what students are taught about any subject necessarily values certain aspects of that subject over others. While this process of selection is inevitable, it is not uniform either across cultures or within a given culture at a particular time (200).

In science, the contextualised nature of scientific knowledge has been discussed by Feyerabend (1978) who argued, for example, that the Copernican revolution happened not because rule and standards which were 'rational' became clear, but rather because of a particular interplay of 'attitudes, discoveries and difficulties' (p.53) that gave Copernicus great importance. An acceptance of the importance of context "acknowledges the indeterminacy of meaning and the importance of individual and communal interpretations and values" (Huot p.558). One tradition within hermeneutics is particularly applicable to this present discussion. That is the perspective of Heidegger and Gadamer which recognises that:

> … the readers preconceptions, "enabling" prejudices, or foreknowledge are inevitable and valuable in interpreting a text. In fact they make understanding possible. Here, the hermeneutic cycle [initial interpretation, validation, revised interpretation] is viewed as including a dialectic between reader and text to develop practically relevant knowledge (Moss 1994 p.7).

In this view, over-commitment to scientific principle and "objectivity" has led to "context-stripping" when making evaluations. But, given the central importance of context, it is hard for raters to agree in "contextually stripped environments" (Huot p.558). Huot describes and interprets a famous study on "reliability" as follows:

> … the most famous study involving the inability of raters to agree on scores for the same papers was conducted by Paul Diederich, John French and Sydell Carlton. Three hundred papers were distributed to 53 readers representing six different fields. Readers were given no sense of where the papers came from or the purpose of the reading. Given this lack of contextual cues, it is not surprising that 90% of the papers got at least seven different scores on a nine-point scale (Huot 1996, p.557).

The recognition of the importance of context is reflected in current work in linguistics about the centrality of context in language use (de Beaugrande and Dressler 1981, Brown and Yule 1983, Halliday1978, Labov 1980, Levinson 1983 all cited in Huot 1996, p.559).

The role of negotiation

In the types of evaluation advocated by Guba and Lincoln, *negotiation* is also central in two ways. They focus on the need for evaluation that is both *responsive* and *constructivist*. In *responsive* evaluation, the evaluator seeks out the views of different stakeholders. In our assessment context, this could include those who teach particular courses, university administrators, students, and employers. We will see this type of movement later in assessment practice where assessors have sought out the views of students (e.g. Darling 2001) and front-line assessors (e.g. Baume and York 2002)[9] in establishing the criteria by which their work should be judged. In this way, the concerns of the different stakeholders may be taken into consideration. It does mean that "traditional experts" and those with positions of power in the assessment process (e.g. some administrators) should give up some control over the assessment process.

In *constructivist* terms, evaluations and assessments are socially contextualised and constructed judgements. Each individual assessor has to reach his or her own contextualised and constructed judgement through an internal process of interpretation. S/he will then enter into a negotiating process with peers. Central is the notion that there is *not* one "ideal", objective grade that we should be striving to reach. On the contrary, Wiggins (1994) has written "all assessment is subjective; the task is to make the judgement defensible and credible" (cited in Allen 1995, p.73). All the assessor can do is to create "a constructed reality that is as informed and sophisticated as it can be made at a particular point in time" (p.44).

As Guba and Lincoln summarise:

> Fourth generation evaluation is a marriage of responsive focusing – using the claims, concerns, and issues of stakeholders as the organising elements – and constructivist methodology – aiming to develop judgemental consensus among stakeholders who earlier held different perhaps conflicting, emic constructions (p.184).

Interpretation and the interpretive community

Reading and writing and assessment in general are viewed as essentially interpretive acts. This ties in with recent poststructuralist conceptions of reading as "a process, a creative interaction between reader and text" (White 1994 p.91). From about the 1930s to the 1950s theories of reading centred around formal criticism which rested on the assumption that meaning existed in the text, independent of the reader or writer. Although poststructuralists such as Bleich, Holland, Fish, Derrida and Barthes take a variety of positions, they all offer one important insight, "opposition to the belief that meaning resides entirely in a text" (White p.93).

> The author's intentions, the reader's individual associations with words, the reading situation, and all kinds of other matters outlawed by formal criticism can now be considered as part of the total meaning a reader creates from the text … In short, these theories of reading have brought a new liberation – some would call it anarchy- into the reading process and placed a much heavier responsibility on the reader to create meanings that may or may not be present on the page for other readers (p.95).

Constructivist, interpretivist assessment does not mean that any interpretation is acceptable, that "anything goes". On the contrary, Moss's (1994) approach to assessment involves a rational debate among a community of interpreters and collaborative inquiry that "encourages challenges and revisions to initial interpretations" (cited in DeRemer 1998, p.27). Stanley Fish articulates the idea of "the interpretive community", rather than suggesting "anything goes" in reading. The interpretive community is made up of those who agree how to read/write texts.

> Interpretive communities are made up of those who share interpretive strategies not for reading (in the conventional sense) but for writing texts, for constituting their properties and assigning their intentions.

[9] These researchers are not (knowingly) working within an interpretivist framework, but are carrying out some activities recommended by them.

> In other words, these strategies exist prior to the act of reading and therefore determine the shape of what is read rather than, as is usually assumed, the other way around … This, then, is the explanation both for the stability of interpretation among different readers (they belong to the same community) and for the regularity with which a single reader will employ different interpretive strategies and thus make different texts (he belongs to different communities) Fish 1980, p.171 cited in White p.99-100).

Various theorists have argued that knowledge is by no means chaotic in that at any one time in any one place there has to be a certain social consensus over what constitutes socially accepted knowledge and extensions thereof (Bakhtin 1986; Bakhtin and Medvedev 1985; Brandt 1992; Feyerabend 1978; Foucault 1980, 1981; Nystrand 1989, 1993). Indeed Foucault, the French polymath and historian, talks about the "order of discourse". These rules govern who is allowed to talk and what they can say.

> In sum, people talk in a context of existing discursive orders that (1) endow people with different qualifications and opportunities to talk, and with different rights to tell the truth; (2) establish regions of knowledge and regions of "silence"; (3) set truth conditions - a "regime of truth"; and (4) link that regime of truth "in a circular relationship with systems of power which produce and sustain it, and to effects of power which it induces and which extend it" (Foucault 1980, p.133 cited in Lindstrom 1992, 104-5).

In the sciences too, Feyerabend has argued that social aspects have a role in deciding what knowledge is valued and accepted at any particular time. In this environment, an "anything goes" scenario is rather unlikely.

Assessors have to try to create an "interpretive community" as described by Fish above. This they can do by discussion of values, agreement on scoring standards, and internalisation of such standards. We will see this type of movement later in assessments where different people involved in assessments have sought discussions about evaluations (e.g. Allen 1995; Broad 2000). Should assessors not be allowed adequate time and other facilities to create this temporary interpretive community, the grading process is likely to break down. White argues that scoring guides should never be imported from other groups. Instead scoring guidelines are devices for building an interpretive community (White p.102).

Guba and Lincoln (1989) describe the hermeneutic, dialectic process involved in constructivist assessment:

> It is *hermeneutic* because it is interpretive in character, and *dialectic* because it represents a comparison and contrast of divergent views with a view to achieving a higher-level synthesis of them all, in the Hegelian sense. Nevertheless, the major purpose of this process is not to justify one's own construction or to attack the weaknesses of the construction offered by others, but to *form a connection* between them that allows their mutual exploration by all parties. The aim of this process is to reach a consensus when that is possible; when it is not possible, the process at the very least exposes and clarifies the several different views and allows the building of an agenda for negotiation (p.149).

The idea of an "interpretive community" maps on to the notion described by Darling (2001) of portfolio construction as a social practice. As she describes a social practice is:

> A complex human activity governed by rules, standards of excellence that are considered in the light of certain virtues and initiated through a particular intention or set of intentions (MacIntyre, 1984). They are carried out within a community in which understandings about certain practices are shared (Benhabib, 1992). Every practice needs to be considered in the light of these features: rules standards, virtues and intentions. These in themselves are complex notions, and when interacting with one another even more so. Together they provide a rich way to view certain kinds of human activity from dressmaking to baseball to composing a musical score, all of which have rules (without which the practices would be unintelligible and in fact, impossible) and standards (you can do these things poorly or well). Importantly, a practice is understood in terms of intentions, intentions that are known to those who engage in the practice and those who have learned to truly appreciate it (p.108).

What would a hermeneutic approach to portfolio assessment involve?

Assessment in a hermeneutic approach would be likely to involve:

- ***holistic, integrative interpretations*** of collected performances that seek to understand the whole in light of its parts. Rather than breaking assessment down into individual components, hermeneutic assessors focus on holistic assessment. As Moss (1994) explains:

  > … most hermeneutic philosophers share a holistic and integrative approach to interpretation of human phenomena that seeks to understand the whole in light of its parts, repeatedly testing interpretations against the available evidence until each of the parts can be accounted for in a coherent interpretation of the whole (p.7).

- ***valuing context-bound knowledge*** about the assessees and assessment environment which it perceives as integral to the production of an assessment artefact (Broad p.248-9). It would privilege readers who are most knowledgeable about the context in which the assessment occurs. Huot (1996) describes a new assessment vision:

  > These new assessment schemes are context-rich and rely upon raters knowing as much as possible about the papers, the students, the purpose of the evaluation, the consequences of their decisions and the decisions of fellow raters. Within these procedures, the sacred cow of writing assessment, interrater reliability, becomes irrelevant because agreement on blind readings is no longer a crucial element for an accurate or valid evaluative decision. Instead, our attention is directed toward creating an assessment environment for reading and writing that is sensitive to the purpose and criteria for successful communication in which student ability in writing becomes part of a community's search for value and meaning within a shared context (p.563).
- ***articulation of the values and judgements of the assessors***, perhaps through discussion of sample texts, other assessment artefacts or entire portfolios. "Articulation" of the values and judgements of the assessors using sample portfolios is a process of drawing out views, rather than imposing from above. Sample texts or portfolios likely to be chosen not as examples of particular qualities of task, but rather to evoke particular evaluative issues or problems (Broad 2000, p.253). This *articulation* is likely to involve discussion around issues important to the assessors rather than a decontextualised scoring rubric. Hermeneutic assessors find that rubrics cannot cover the range of portfolios they are presented with; that it is hard to find portfolios or even single pieces of writing (or other assessment artefacts) that conform to the ladder of quality outlined in most scoring rubrics; rubrics can be misleading in that they support the notion that assessment is a transparent, neutral process. In his suggestions for assessment of portfolios in a composition programme in an American university, Broad advocated not having a scoring rubric as such rubrics do not cover the range and diversity of portfolio offerings and do not acknowledge the value and inevitability of the individual assessor's particular interpretation. He suggested that assessors discuss portfolios and that the intermingling and interactions of their views will provide a richer interpretation of the texts read. Broad wrote that:
- Hermeneutic assessment does not pretend to guarantee that various evaluators will produce identical decisions; instead, it provides a multiperspectival, dialogic exchange by which participants may raise questions and challenges regarding their peers' decisions. What emerges is a more dynamic process of evaluation and one more true to participants rhetorical and pedagogical principles (Broad 2000, p.249).
- ***grounding*** those ***interpretations*** in

  a) available textual and contextual evidence and also in

  b) a rational debate among the community of interpreters (Moss 1994, p.7).

  Interpretations would be accountable to others through a process of critical discussion (Broad p.249). A variety of interpretations is not seen as a breakdown of the assessment process; the assessment process requires discussions about the different interpretations which may shift in the light of the discussion.
- perhaps but not necessarily pass/fail grading rather than numerical grades in that interpretivist assessment perceives the allocation of grades to complex texts as an impossible and misleading task (Broad 2000 p.251).
- ***judgement by experts in the field.*** In terms of Darling's conception of social practice, those who are themselves initiates are the ones who will be in a position to make judgements. They are the ones who understand the rules, standards and intentions behind the process and products of the social practice. "One needs to be sufficiently 'inside the practice' to fully know what the standards of excellence are, and getting inside the practice takes time and self-discipline as well as instruction or apprenticeship" (p.109). As MacIntyre describes of the process of judgement:

  > A practice involves standards of excellence and obedience to rules as well as the achievement of goods. To enter into a practice is to accept the authority of standards and the inadequacy of my own performance as judged by them. It is to subject my own attitudes, choices, preferences and tastes to the standards, which currently and partially define the practice (cited in Darling 2001, p.109).

Such a mode of assessment is based on acknowledging and valuing differences of judgements about texts brought for assessment.

Conditions required for interpretive assessment to work

A hermeneutic approach to assessment requires the fulfilment of several conditions if it is to work. As Guba and Lincoln (1989) outline, it requires:

(1) A commitment from all parties to work from a *position of integrity.* There will be no deliberate attempt to lie, deceive, mislead, hide, or otherwise offer misconstructions. …

(2) *Minimal competence* on the part of all parties to *communicate*. Holders of different constructions must be able to offer their own constructions, and to offer criticisms of the

constructions of others, meaningfully. [So very young children, those who are severely mentally handicapped, psychotics would be excluded].

(3) A willingness on the part of all parties *to share power.*

(4) A willingness on the part of all parties *to change* if they find the negotiations persuasive. Those who are 'true believers' in some construction, or who fear 'revisionists,' de facto cannot carry out a meaningful negotiation.

(5) A willingness on the part of all parties to *reconsider their value positions* as appropriate. This condition shares the same caveats as the preceding item.

(6) A willingness on the part of all parties to *make the commitments of time and energy* that may be required in the process (p.149-150).

Standards and criteria for judging the adequacy of constructivist assessment

In the positivist psychometrically-based approach to assessment, validity and reliability are used to judge the adequacy of an assessment. These depend on the notion that objective standards can be set and that an objective reality exists. This is clearly unsuitable for constructivist assessment. Mishler (1990) argued that:

> Reformulating validation as the social discourse through which trustworthiness is established elides such familiar shibboleths as reliability, falsifiability, and objectivity. These criteria are neither trivial nor irrelevant, but they must be understood as particular ways of warranting validity claims rather than as universal, abstract guarantors of truth. They are rhetorical strategies … that fit only one model of science (p.420 qtd in Moss p.10).

Guba and Lincoln (1989) suggest alternative criteria and standards of judging the adequacy of assessment. They focus mainly on criteria and standards for judging institutional evaluations. However, many of the points they make can be adapted to assessment of individuals.

Parallel criteria

Guba and Lincoln focus first on what they call *parallel* criteria for assessment: credibility, transferability, and dependability. Here they seek to provide parallel criteria with those criteria used in positivist, psychometrically based assessment. They suggest that *credibility* in constructivist assessment is parallel to internal validity in positivist approaches to assessment. They suggest that internal validity in positivist assessment is concerned with establishing the isomorphism between the findings of the assessment and an objective reality. They suggest that in constructivist assessment the focus shifts to establishing a match between the assessment and the constructed realities of those being assessed, presumably other assessors and other stakeholders (in our case employers, academic staff, other interested individuals). In order to increase the likelihood of this match, they suggest a range of techniques, traditionally used by anthropologists, sociologists and educational researchers etc. These techniques include:

- prolonged engagement in the assessment process;
- persistent and detailed observation of the assessment situation;
- peer debriefing;
- negative case analysis;
- progressive subjectivity; and
- member checks (p.236-9).

In the context of our type of assessment, this tends to suggest methods of assessment such as portfolios (rather than one-off examinations) where work can be viewed over a period of time. Indeed, Moss (1994) suggested that multiple and varied sources of evidence was a criterion for justifying the interpretation. Should the course teacher him/herself be assessing the course, the involvement of a disinterested peer with whom the assessor could discuss at least some of the assessment tasks and decisions would be recommended. Negative case analysis in the case of portfolios would involve looking at unexpectedly strong or weak pieces of work to determine whether the assessment verdict should be altered. In order to achieve progressive subjectivity, the assessor should keep some kind of a record of assessment decisions to make sure that assessment insights develop and do not remain locked into initial impressions. In the case of assessment of individuals, member checks would probably involve checks with stakeholders such as students to make sure that the assessments tallied with their own perceptions. It could also involve checking with those where the students go on to work or study to make sure that the students have been adequately assessed for the purposes required. Clearly this last requirement would have to be addressed at a level beyond that of the individual assessor.

Guba and Lincoln suggest that *transferability* is parallel to external validity or generalisability in positivist assessment. They suggest than in order to achieve transferability, a major technique is "thick description" (Geertz 1973), that is to provide "an extensive and careful description of the time, the place, the context, the culture in which those hypotheses were found to be salient" (p.241-2). In individual assessment, presumably this could take the form of a course description, assessment criteria, and a written report if numbers of

students are sufficiently low and staff resources adequate to deal with this. The idea of profiles (see Chapter Two, *Profiling* p.19-20) might be relevant here.

Guba and Lincoln suggest that *dependability* is parallel to reliability in positivist assessment. Constructivist assessment views changes in constructions as normal and desirable as the assessment process moves forward and as constructions become increasingly sophisticated. They suggest, however, that people external to the assessment process should be able to track shifts in construction and assessment decisions. Presumably, this would entail some kind of documentation of assessment decisions.

They suggest that the major check on the quality of the constructivist assessment process is within the nature of the process itself:

> The opportunities for error to go undetected and/or unchallenged are very small in such a process. It is the immediate and continuing interplay of information that militates against the possibility of noncredible outcomes. It is difficult to maintain false fronts, or support deliberate deception when information is subject to continuous and multiple challenges from a variety of stakeholders. The publicly inspectable and inspected nature of the hermeneutic process itself prevents much of the kinds of secrecy and information poverty that have characterized client-focused evaluations of other generations [of evaluation] (p.244).
>
> Further, the possibility that the so-called biases or prejudices of the evaluator can shape the results is virtually zero, provided only that the evaluation is conducted in accordance with hermeneutic dialectic principles. (The argument that not all evaluators will 'play the game' honourably and honestly is unconvincing. The same observation can be made of inquiry conducted within *any* paradigm, as recent experience so well attests.) So long as the evaluator's constructions (to which she or he is entitled as is any other constructor; calling them biases may have persuasive value but is hardly compelling) are laid on the table along with all the others and are made to withstand the same barrage of challenge, criticism, and counterexample as any others, there is no basis for according them any special influence, for better or worse (p.244).

Authenticity criteria

Guba and Lincoln are not satisfied with these criteria which, as they point out, have their roots in the demands of positivist assumptions and which concern mainly methodological issues. They therefore suggest a range of what they call *authenticity criteria.* These derive directly from constructivist assumptions. These were drawn up to describe evaluations rather than assessments so some adjustments will be necessary, but the underlying principles remain the same. Openness and negotiation are the keystones.

They suggest that the evaluation/assessment has to be *fair*. Broadly this is to be achieved through openness about the process and negotiation. Broad (2000) and Moss (1994) also say that the assessor's interpretation would be justified by criteria such as the transparency of the trail of evidence leading to the interpretations, which allows users to evaluate the conclusions for themselves" (Moss 1994, p.7). Discussions about assessment would be documented and publicised (Broad p.253). Criteria for assessment should be very clear. In the last resort, there should also be some kind of appeal process, should someone feel that rules have not been observed.

Potential threats to fairness can be imagined. People will have strong feelings about the grading of students, especially their own, and proper safeguards have to be in place to ensure against various dangers such as:

- a person in a powerful department position exerting pressure on colleagues in assessment groups to favour certain students
- more articulate, determined members of the group overriding others.

Guba and Lincoln also suggest *ontological authenticity*, that "individual respondents' own emic constructions are improved, matured, expanded, and elaborated, in that they now possess more information and have become more sophisticated in its use" (p.248). In the context of assessment, presumably this would mean adequate feedback mechanisms and explanations of the assessment with presumably formative assessment before a summative assessment.

They argue that:

> To substitute relativity for certainty, empowerment for control, local understandings for generalized explanation, and humility for arrogance, seems to be a series of clear gains for the fourth generation evaluator (p.48).

Not all may agree with this.

Conclusions

We have sought to present the two sides of the theoretical underpinnings in the assessment debate. Neither approach guarantees fairness, agreement among raters or any other aspect of assessment. However, as Moss (1994) argues, it is important to lay the assumptions and consequences of each approach on the table so that those involved in assessment can make better informed choices. Moss does not advocate, for example:

> The abandonment of reliability. Rather I am advocating that we consider it one alternative for serving important epistemological and ethical purposes- an alternative that should always be justified in critical dialogue and in confrontation with other possible means of warranting knowledge claims. As Messick (1989) has advised, such confrontations between epistemologies illuminate assumptions, consequences, and the values implied therein. Ultimately, the purpose of educational assessment is to improve teaching and learning. If reliability is put up on the table for discussion, it if becomes an option rather than a requirement, then the possibilities for designing assessment and accountability systems that reflect a full range of valued educational goals become greatly expanded (p.10).

Given the huge intellectual challenges to positivism in many disciplines in recent decades, it is astonishing in one sense that the debates about assessment are so dominated by positivist approaches. In another sense, given the societal and political pressures on educationalists mentioned earlier, it is not in the least surprising.

Figure 2 summarises some of the major feature of positivist and interpretivist assessment. The first part of the table is compiled from information in Guba and Lincoln 1989. The second part of the table is drawn from Moss (1994), Huot (1996) and Broad (2000).

Figure 2 Features of positivist and interpretivist assessment

	Positivist assessment	**Interpretivist assessment**
Ontology	"A realist ontology asserts that there exists a single reality that is independent of any observer's interest in it and which operates according to immutable natural laws, many of which take cause-effect form. Truth is defined as that set of statements that is isomorphic to reality" (p.84)	"A relativist ontology asserts that there exist multiple, socially constructed realities ungoverned by any natural laws, causal or otherwise. "Truth" is defined as the best informed (amount and quality of information) and most sophisticated (power with which the information is understood and used) construction on which there is consensus (although there may be several constructions extant that simultaneously meet the criterion" (p.84).
Epistemology	"A dualist objectivist epistemology asserts that it is possible (indeed, mandatory) for an observer to exteriorize the phenomenon studied, remaining detached and distant from it (a state often called 'subject-object dualism') and excluding any value considerations from influencing it (p.84).	"A monistic, subjectivist epistemology asserts that an enquirer and the inquired-into are interlocked in such a way that the findings of an investigation are the *literal creation* of the inquiry process. Note that this posture effectively destroys the classical ontology-epistemology distinction" (p.84)
Nature of truth	"The truth of any proposition (its factual quality) can be determined by testing it empirically in the natural world. Any proposition that has withstood such a test is true: such truth is absolute" (p.104).	"The truth of any proposition (its credibility) can be determined by submitting it semiotically to the judgement of a group of informed and sophisticated holders of what may be different constructions. Any proposition that has achieved consensus through such a test is regarded as "true" and reconstructed in the light of more information or increased sophistication; any truth is relative" (p.104).
Limits of truth	"A proposition that has not been empirically tested cannot be known to be true. Likewise, a proposition incapable of empirical test can never be confirmed to be true" (p.104).	"A proposition is neither tested nor untested. It can only be known to be 'true' (credible) in relation to and in terms of informed and sophisticated construction" (p.104).
Nature of evaluation[10]	"Evaluation is a form of scientific inquiry and hence has all attributes of that genre" (p.109).	"Evaluation is a form of constructivist inquiry and hence has all the attributes of that genre" (p.109).
Values and evaluation	"Evaluation produces data untainted by values. Values are intrusive to the evaluation process and distort scientific data by, for example, biasing them" (p.109).	"Evaluation produces reconstruction in which 'facts' and 'values' are inextricably linked. Valuing is an intrinsic part of the evaluation process, providing the basis for attributed meaning" (p.109).
Objectivity of evaluation findings	"Evaluators can find a place to stand that will support the objective pursuit of evaluation activities" (p.110).	"Evaluators are subjective partners with stakeholders in the literal creation of evaluation data" (p.110).
Function of evaluation	"Evaluators are the communication channels through which literally true data are passed to the audience of evaluation reports" (p.110)	"Evaluators are orchestraters of a negotiation process that aims to culminate in consensus on better informed and more sophisticated constructions" (p.110).

[10] The term "evaluation" in the first sections of the table comes from an American book, Guba and Lincoln 1989. Then tend to focus on evaluation in the British sense of the word, but much of what they are saying is also applicable to assessment in the British sense.

	Positivist assessment	Interpretivist assessment
Methodology	"An interventionist methodology strips context of its contaminating (confounding) influences (variables) so that the inquiry can converge on truth and explain nature as it really is and really works, leading to the capability to predict and to control" (p.84).	"A hermeneutic methodology involves a continuing dialectic of iteration, analysis, critique, reiteration, reanalysis, and so on, leading to the emergence of a joint (among all inquirers and respondents, or among etic or emic views) construction of a case" (p.84).

Compiled from information in Guba and Lincoln 1989.

Scoring guidelines	Raters are given scoring rubrics which set out features of writing quality. The assumption is that "writing quality can be defined and determined" (Huot p.551).	Understandings and agreements are developed within the community of raters which are suitable for the local context
Training of raters	Careful training enables trainers to agree on allocation of scores to work of a particular quality. The assumption is that such agreement is possible and desirable.	"Articulation" of the values and judgements of the assessors using sample portfolios, a process of drawing out views, rather than imposing from above. Sample texts or portfolios likely to be chosen not as examples of particular qualities of task, but rather to evoke particular evaluative issues or problems (Broad 2000, p.253)
Scoring process	• Individual tasks tend to be scored separately by raters (e.g. in portfolio research it would probably be recommended to score one piece of work at a time across a range of portfolios • Tasks are scored numerically.	• Multiple tasks tend to be assessed together in order to give a holistic, integrated impression of the student's work • Tasks are less likely to be scored numerically
	Raters have no knowledge of the judgements of other raters. Each rater allocates a grade which is then fixed.	Raters engage in discussion with one another about the student's work in order to have a "rational debate among the community of interpreters" about the quality and nature of the work. Initial interpretations may shift as a consequence of discussion.
	Raters have no additional knowledge, apart from task they are rating, about student	Raters' interpretations are grounded in textual and contextual information available about the student. Raters who have the most knowledge about the student are privileged.
	• Achievement is assessed by aggregating scores from independent raters and tasks • Scores are referenced to specified criteria or norms • These scores often come with interpretative guidelines which are validated by previous research	"The interpretation might be warranted by criteria like a reader's extensive knowledge of the learning context; multiple and varied sources of evidence; an ethic of disciplined, collaborative inquiry that encourages challenges and revisions to initial interpretations; and the transparency of the trail of evidence leading to the interpretations, which allows users to evaluate the conclusions for themselves" (Moss 1994, p.7)
Generalisability (in both positivist, psychometrically-based and interpretivist approaches, assessment involves generalising from the observable tasks in the assessment to performance on	• Consistency across independent tasks and readers warrants generalisability in the shape of quantitative scores. Consistency and standardisation should be maintained across time, location and rater. If tasks and raters are not consistent, then the validity of the assessment is in doubt.	• Consensus among the interpreters warrants the validity of the interpretation • Interpretation is reached by discussion among the readers which continues while different interpretations are explored and integrated into a holistic understanding. A well-documented report describes to others how the interpretation has been reached. • The interpretation may become the basis for pedagogical action, the success of which may further validate the interpretation.

other unobservable and unobserved tasks)		
Fairness	• Objective, detached decisions in high-stakes assessment are seen as an essential pre-requisite for fairness to students involved in the assessment and for those who use the results (such as employers) • Consistency across tasks may be unfair in that students may be differentially familiar with the task they are asked to do.	• Consistency among tasks is not necessary or even desirable. What is more desirable is to allow selection of tasks to be assessed by the students in that they can choose the tasks which best represent their strengths. • This may be unfair if students are not prepared for such a selection so teachers would have to prepare them adequately for this task. • Discussion and multiple interpretations necessary for high stakes decisions on assessment. Tends to view the "objective" and non-contextualised assessment of psychometrically-based approaches as authoritarian and potentially arbitrary as it "silences" the voices of those who are most knowledgeable about the context and those most directly affected by the results" (Moss p.10).

Compiled from information in Moss (1994), Huot (1996) and Broad (2000).

Having explored the theoretical approaches to assessment, with their underlying assumptions and different practical implications, we will now turn to a review of the research on agreement over outcomes in portfolio summative assessment.

A review of the research on agreement over outcomes in portfolio summative assessment

This discussion will focus only on issues relating to agreement over portfolio assessment outcomes in summative (high stakes) assessment, since agreement in formative (low-stakes) assessment is not usually a major issue. Agreement over grades in summative assessment has been a major concern in portfolio assessment, given the complex nature of the assessment task (multiple and probably complex assignments) and the nature of the context in which the assessment takes place.

The discussion over portfolio assessment takes places in the social, political, financial and educational context described previously *A basic assessment dilemma in UK higher education today* p.35. There is enormous pressure from outside universities to have systems of assessment which are "reliable", cost-effective, efficient, fair, usable by educators, useful to employers and so on (*Editorial* p.306). At the same time, educationalists are aware of the need to develop assessment systems which are aligned with current thinking on learning theory.

The discussion also takes place in a context which is philosophical, theoretical, but also profoundly practical. In terms of the philosophical and theoretical context, researchers oriented towards positivism assume that it should be possible to reach one ideal, objective assessment of a portfolio through appropriate training of assessors, construction of clear guidelines and other such measures. Researchers oriented towards interpretive, constructivist assessment assume that different perspectives will be brought to bear when assessing a portfolio and that these will result in differing interpretations and assessments which should be valued. Agreement in interpretivist assessment involves debate among the rational, interpretive community of interpreters/graders. See the previous section *Theoretical dilemmas and practical implications: positivist and interpretivist approaches to summative assessment p.39-52* for a fuller explanation of these differences. We will see how these theoretical approaches were realised in practice in some of the studies that follow.

This philosophical and theoretical division, although basic and important, is perhaps less stark than the binary division above might suggest. Firstly, many of those who operate in the traditional, positivist, psychometrically-based mode of assessment are aware of the need for an extended conception of validity as developed in recent years in portfolio assessment, and indeed in any assessment. They are aware of the

dangers of achieving reliability through measuring only trivial aspects of work. For example, Baume and Yorke (2002) describe the care taken in an Open University course for accrediting university lecturers to assess a course "validly" by constructing clear guidelines for assessors which reflect course aims and practice and by making the criteria for assessment explicit to assessees and tutors. LeMahieu, Gitomer and Eresh (1995) talk about the need to avoid achieving reliability through "simplification, even trivialization, of the judgements sought. To the extent that subtlety, insight, even wisdom, are pursued in the judgement exercised, then the source of clarity, consistency, and discipline in those evaluations must be found elsewhere" (p.25). Secondly, some are edging towards acknowledging the need for raters to have common understandings as well as a knowledge of the context in which they are assessing and to discuss assessment with other raters. In a clear-sighted article, LeMahieu, Gitomer and Eresh (1995), in their discussion of school level portfolio assessment in Pittsburgh (USA), stress the importance of developing common understandings not only at the level of developing the scoring rubric, but also at the much earlier stage of curriculum development in order to ensure that classroom aims, curriculum aims and assessment are consistent. In reviewing the evidence on interrater reliability, Baume and Yorke have suggested that it is more difficult for raters to make judgements when there is insufficient information about the assessee (e.g. other work not in the portfolio, context in which the work was produced). Supovitz, MacGowan and Slattery (1997), in their study of assessment of primary school portfolios in New York also make this point. This ties in with the comments of interpretivist assessors about the need to understand the context of an assessment before making a judgement. However, for assessors in the positivist tradition, the discussion about rating precedes assessment instead of also continuing at a later stage. Moreover, positivist, psychometrically oriented researchers still operate on the basic assumption that there is an objective ideal grade out there which is to be discovered rather than one which is agreed on upon by those located in a particular social and educational context.

Much of the debate about agreement over grades in portfolio assessment has been couched in terms of the mainstream positivist, reliability framework. This paper will first discuss this research before turning to agreement over grading in hermeneutic assessment of portfolios.

Positivist researchers on portfolio summative assessment have been interested in reliability from various points of view (e.g. intrarater reliability, interrater reliability, how people actually use scoring rubrics). However, most of the research seems to address issues of interrater reliability so this discussion will focus largely on that with some discussion of how people actually use scoring rubrics. In this section on *positivist* approaches to portfolio summative assessment, we will (1) introduce the topic from the positivist point of view; (2) review specific studies; (3) look at suggestions for increasing interrater agreement; and (4) point out gaps in the research on interrater reliability. In the section on research on *interpretivist* approaches to portfolio summative assessment, we will (1) review specific articles; (2) point out gaps in the research. The final section on agreement in portfolio summative assessment draws out particular points of interest from the reviews.

Figure 3 below provides an overview of issues that the articles address. It indicates issues that each article investigates. Figure 4 below indicates the sector and specific subject focus of each article as well as the country in which the research was carried out. These articles are largely, although not all, empirical research articles. Should the reader wish to pursue particular issues, consulting Figures 3 and 4 should allow him/her to locate which summaries are likely to be particularly relevant?

Figure 3 Agreement over rating: summary of issues that each article examines

Issue examined	Article
Varying rates of "reliability" (positivist approach)	Baume and Yorke (2002); Centra (1994); Davis et al. (2001); Herman and Winter (1994); Supovitz, MacGowan and Slattery (1997)
Role of contextual information on grading	Allen (1995); Broad (2000); Supovitz, MacGowan and Slattery (1997)
Effects of the inclusion of the class teacher in the grading process	Allen 1995; Broad (2000); Supovitz, MacGowan and Slattery (1997)
Effects of inclusion of other people known to the assessee in the grading process	Centra (1994)
Training and preparation of raters (positivist and interpretivist)	Allen (1995); Baume and Yorke (2002); Broad (2000); LeMahieu, Gitomer and Eresh (1995)
Fairness	Baume and Yorke (2002); Broad 2000; Nystrand, Cohen and Dowling (1993)
Particular elements of an assessment causing problems	Baume and York (2002); Centra (1994)
Consistency of portfolio grading with other elements on an examination	Davis et al (2001)
Portfolios rated holistically	Allen (1995); Broad (2000); Nystrand, Cohen and Dowling (1993)
Portfolios rated by particular elements	Baume and Yorke (2002)
Portfolios rated by individual assignment	Nystrand, Cohen and Dowling (1993)
Integration of portfolios into overall curriculum	LeMahieu, Gitomer and Eresh (1995)
Effects of the method of calculating the scores on the final grade	Allen (1995); Baume and York (2002); Fourali (1997)
Speed of rating	Nystrand, Cohen and Dowling (1993)
Suggestions for increasing interrater reliability (positivist approach)	Baume and Yorke (2002); Centra (1994); Davis et al. (2001); Nystrand, Cohen and Dowling (1993); Reckase (1995)
Suggestions for arriving at agreement on grades (interpretivist approach)	Allen (1995); Broad (2000)
How graders actually use scoring rubrics	DeRemer (1998); Huot (1993); Pula and Huot (1993); Smith (1993); Vaughan (1991); Wolfe and Feltovitz (1994); Wolfe and Ranney (1996)
Professional growth involved in discussing portfolio assessment within the assessment process	Allen 1995, Supovitz, MacGowan and Slattery (1997)

Figure 4 Agreement over rating: sector and specific area focus of articles

Educational sector	Specific area (if relevant) and country	Article
Schools	Various subjects (US)	Herman and Winters (1994)
	Language arts (US)	Supovitz, MacGowan and Slattery (1997)
Higher education	University teacher training (UK)	Baume and Yorke (2002)
	Faculty assessment for contract renewal and promotion (US)	Centra (1994)
	Undergraduate medical education (UK)	Davis et al (2001)
	Various subject areas	Nystrand, Cohen and Dowling (1993)
	Not specific, abstract	Reckase (1995)
	Freshman Writing	Broad (2000)
	Various	Allen (1995)

Interrater "reliability" in the positivist, psychometrically-based tradition

Herman and Winters (1994) claim that "interrater agreement is accepted as the foundation upon which all other decisions about portfolio quality are made" (p.49). In terms of portfolios, reliability of assessment entails the following according to Herman and Winters (1994):

> Raters who judge student performance must agree regarding what scores should be assigned to students' work within the limits of what experts call "measurement error." Do raters agree on how a portfolio ought to be scored? Do they assign the same or nearly similar scores to a particular student's work? If the answer is no, then student scores are a measure of who does the scoring rather than the quality of the work. Interrater agreement is accepted as the foundation upon which all other decisions about portfolio quality are made (p.49)

They argue further that reliability requires "score stability for the same student on different occasions; score stability of papers/entries given different contexts or 'portfolio sets' in which a portfolio is scored; score consistency across 'like' tasks" (p.50).

The general position of positivist, psychometrically-based researchers is that interrater reliability can be quite high if the assessment assignment is relatively straightforward (which is not usually the case with portfolios), but becomes increasingly difficult as the assessment assignment increases in the complexity necessary for greater authenticity (e.g. Miller and Legg 1993; LeMahieu, Gitomer and Eresh 1995). Portfolio assessment has, therefore, traditionally presented considerable problems to those for whom "reliability" is a major issue.

However, typically, assessors working in this tradition have at times achieved quite high standards of agreement between raters. In the quest to achieve reliability in assessing portfolios, they have used scoring rubrics and detailed training in an effort to provide a standard by which writing should be judged.

Various studies have been conducted where the scores of two or more raters, operating independently, have been compared and percentages of (dis)agreement over number of portfolio calculated. Researchers have documented widely varying ranges of interrater agreement. Researchers have commented on factors which increase and decrease "reliability".

Reviews of specific studies

Herman, Joan, and Lynn Winters. "Portfolio Research: A Slim Collection." ***Educational Leadership*** **(1994): 48-55.**
LeMahieu, Paul G., Drew H. Gitomer, and JoAnne Eresh. "Portfolios in Large-Scale Assessment: Difficult But Not Impossible." ***Educational Measurement: Issues and Practice*** **(1995): 11-28.**
Heller, Joan I., Karen Sheingold, and Carol M. Myford. "Reasoning About Evidence in Portfolios: Cognitive Foundational for Valid and Reliable Assessment." ***Educational Assessment*** **(1998): 5-40.**
In their 1994 review of portfolio research in schools in the United States, where significant amounts of research on portfolio research have been done, Herman and Winter found varying levels of interrater reliability in different studies. They cited a study by Koretz, Stecher and Deibert (1993) on mathematics and writing portfolio assessment in Vermont schools which reported interrater reliability at .28 or .60, depending on how the scores were aggregated. Another study by LeMahieu et al. (1993) on writing portfolios in

Pittsburgh found far higher correlations, .60 to .70, despite the raters having considerable latitude "in selecting pieces to rate and the broad scope of the scoring criteria" (cited in Herman and Winter 1994, p.50). LeMahieu, Gitomer and Eresh (1995) suggest various reasons for this relatively high correlation: a smaller number of more highly trained graders than Koretz used, careful integration of portfolio assessment into the overall curriculum, and development of the scoring rubric as a shared interpretive framework. In another study on elementary schools portfolios, Herman et al. (1993) found average correlations between raters of .82 (cited in Herman and Winter 1994). In studies of portfolios produced in schools, Heller Sheingold and Myford (1998) found that percentage of rater agreement on assessment of portfolios was more than 90% if raters were allowed to be more than one point apart.

Supovitz, Jonathan A., Andrew MacGowan III, and Jean Slattery. "Assessing Agreement: An Examination of the Interrater Reliability of Portfolio Assessment in Rochester, New York." ***Educational Assessment*** **4 (1997): 237-259.**

Supovitz, MacGowan and Slattery (1997) conducted research on portfolio assessment in primary schools in New York. The purposes of these portfolios are "to provide information to classroom teachers about their students' progress and to give teachers ongoing feedback so they can adjust their instruction. The local authority sponsored a summer institute to investigate its system of portfolio assessment. 20 teachers from the local area who were familiar with the portfolio system were brought together to rate 400 portfolios selected randomly. The portfolios contained a selection of student writing prescribed by the local authority. There was also a selection of reading portfolios which will not be discussed here. The teachers were given a five stage development model on which to assess the portfolios. The classroom teachers had already rated the portfolios. The institute aimed to bring together people who were familiar with the portfolio system, but who did not know the individual students.

The external teachers found it hard to rate 10% of the writing portfolios. These difficulties arose because of insufficient evidence in the portfolio. It was somewhat difficult to make the connections between the work in the portfolio and the developmental stage grading scale. Sometimes the external raters could not tell whether a piece of writing was initiated by the child or copied from the blackboard because of insufficient labelling.

Despite these problems, the researchers found reliability coefficients ranging between .68 for kindergarten ratings of writing portfolios to .73 for Grade 1 listings. The researchers concluded that this level of reliability was too low for high stakes accountability and was associated with lack of evidence in the portfolio. This could arise because of difficulties in making links between the developmental levels in the grading scheme and the actual work in the portfolio; inadequate requirements from the local authority; and teachers not including all required pieces in the portfolio because of lack of perceived relevance to the children's need or lack of time. The external teachers, who were familiar with the classroom problems of implementing the portfolios, were able to explain that it would have taken far too much time to include all the work required in sufficient detail for external raters to be able to rate a portfolio. The classroom teacher responsible for the portfolios compilation in her class, however, would have had sufficient context to be able to rate many more of the portfolios and to make the links between the developmental level of the child and the grading scheme.

The researchers investigated whether the classroom teachers systematically gave higher marks than the external teacher raters. They found not for the writing portfolios and that any difference that did exist arose out of the extra knowledge the classroom teachers had. The researchers concluded that "classroom teachers should be considered a valuable resources for scoring assessments used for accountability purposes" (p.257). They also reported that the summer institute teachers commented that the experience of discussing student work with their peers gave them a better understanding of how children learned and how to distinguish more clearly between different levels of the developmental scheme provided for grading. It is not clear whether classroom teachers currently engage in this discussion as part of the assessment process, but the research findings seem to indicate that would be a good idea (p.257).

Baume, David, and Mantz Yorke. "The Reliability of Assessment by Portfolio on a Course to Develop and Accredit Teachers in Higher Education." ***Studies in Higher Education*** **(2002).**

Research has also been done on interrater reliability in higher education contexts. For example, Baume and Yorke (2002) found that in assessing individual elements of a teacher accreditation course at the Open University (and this portfolio assessment required 75 marks for each portfolio), there were high rates of "close" if not "exact" agreement – 85% or above. However, the agreement rate for passing or failing for the overall portfolio was only 60% because the somewhat complex system of computing the scores had brought down the overall level of agreement. In this study, assessment was especially problematic in that assessors were assessing a range of outcomes and underpinning values. They were not assessing individual pieces of work, but instead looking for a complex range of attributes across different tasks.

Baume and Yorke found that particular areas provoked more disagreement than others. The difficult areas were where assessees were required (1) "to review their teaching, analyse their teacher development needs and make a plan for their continuing professional development in order to address the needs

identified"; (2) keep "records of teaching support and academic administrative work"; (3) present evidence of "concern for equal opportunity" and (4) reflection. Moreover, within (1), it was the second two elements that caused more discrepancies than the first. Baume and Yorke say that these areas where there is greater disagreement suggest that there are differences in perspective about what is to be valued in a portfolio and "hints therefore at problems with validity". A hermeneutic research would, of course, argue that differences in perspective are valuable.

The response from both approaches to assessment (positivist and interpretivist) to the problem may not be so different in practice. Baume and Yorke report that "the course team now gives greater emphasis in assessor briefings to identifying and judging needs analyses and plans for continuing professional development". If this emphasis includes discussions about the views of the assessors, the inclusion of their perspectives in actions taken and the reaching of a consensus about what is reasonable and to be valued, this would not be a million miles from what a hermeneutic assessor would ask for. A hermeneutic assessor though would ask, in addition, for discussions following on from the assessors reading individual portfolios to reach further consensus.

Assessment was further complicated in that a course team produced and ran the course (and moderated the assessment process) while the tutors taught and assessed the course. Many of the people involved in the course were geographically distant from one another. The process of teaching and assessing this course is, therefore, complex. There are various points at which the many people involved communicate, but the straightforward group discussions about individual cases advocated by the hermeneutic assessors might be somewhat logistically complicated in this case.

The study by Baume and Yorke also raises the issue of fairness in that all assessees were asked to present evidence for point (3) above, but clearly some were working in educational environments where opportunities to demonstrate concern for and action concerning equal opportunities were more easily available.

Centra, John A. "The Use of the Teaching Portfolio and Student Evaluations for Summative Evaluation." *Journal of Higher Education* 65 (1994): 555-570.

In another study by Centra (1994), portfolios had been used in a North American higher education context to evaluate the teaching effectiveness of each faculty member of one community college for purposes of contract renewal and promotion.

Like Baume and Yorke, Centra was able to locate the particular elements of the portfolio assessment where there was greater or lesser agreement between raters. He found greater variability in assessment of "teaching" and "service" than on assessment of "credentials" and "participation in professional associations". This is hardly surprising given that the second two elements are more factual. In the case of teaching, for example, the assessors may have had a limited context for assessing the faculty member. Apart from the dean, who made at least one visit to each classroom, the other assessors were dependent on the content of the portfolio for the grade they allocated.

Each portfolio had been evaluated on several elements on an externally constructed and criterion-referenced system by three people: a peer chosen by the faculty member; a dean; and a peer chosen by the dean. The study found that ratings by the peer chosen by the faculty member did not correlate significantly with either the dean's ratings or those of the other peer. The peer chosen by the faculty member scored higher than the other two people. Here we can see the difficulties of achieving agreement in the grading of a portfolio where one assessor is likely to be predisposed to grade higher than the others. In one sense, this is not a problem here, since presumably the "bias" worked in one direction and by the time the three marks were averaged out, a reasonable grade might be reached. A hermeneutic assessor would probably argue that were the three assessors to sit down and discuss their grades, consensus could be reached as the high graders would have to justify their grades and in doing so would either persuade the other graders or themselves be persuaded to moderate their marks.

Davis, M. H., et al. "Portfolio Assessment in Medical Students' Final Examinations." *Medical Teacher* 23 (2001): 357-365.

Davis et al. (2001) carried out research on portfolio assessment in undergraduate medical education in Dundee. Portfolio assessment for part of the final examinations was just being introduced. They found low/moderate agreement between ratings given on portfolio research and on other aspects of the final examination. With the extended multiple choice examination (EMI) the correlation was 0.42; with the constructed response question (CRQ) paper which assessed higher-order thinking skills such as problem solving as well as knowledge the correlation was 0.42; with the (OSCE)[11] 0.47. Rather than concluding the portfolio research was inherently unreliable, the researchers concluded that portfolio assessment was "measuring both common abilities as well as abilities that are different from those tested in the other

[11] i.e. Objective Structured Clinical Examination. An OSCE consists of various "stations" , at each of which and usually within a few minutes only the candidate must perform a standardised clinical task.

examinations" (p.363). The students, who were given a questionnaire, felt that the examiners applied different standards in assessing different portfolios.

In this case, it seems that the research is rather inconclusive. We are not presented with evidence which would allow us to make a judgement about whether the discrepancies in grades between the portfolio assessments and the other grades is because the portfolios are examining different abilities or because the grading of the portfolios was basically flawed. This is essentially an account of the introduction of portfolio assessment in a course and it makes clear that those responsible will learn from this first introduction of portfolios and refine the system of grading for the following year. At that point, were research to be done, it would be possible to see whether there were greater consistency between portfolio grades and those of other elements of the examination. It would also be interesting and useful to know whether there were a strong correlation between the grades on the other elements of the examination.

Nystrand, Martin, Allan S. Cohen, and Norca M. Dowling. "Addressing Reliability Problems in the Portfolio Assessment of College Writing." *Educational Assessment* 1 (1993): 53-70.

Nystrand, Cohen and Dowling (1993) in their study of portfolio rating in the University of Wisconsin focused on the differences in interrater reliability according to how the portfolio was read and rated. Each portfolio "contained every piece of writing- course papers, exams, lab reports, and reviews on various topics- that each student wrote as part of coursework" (p.54). The students were drawn from various subject areas. The portfolios were rated on the "degree of reflection and extent of text elaboration" (p.53). An adapted form of Britton et al.'s (1975) scale with 8 levels was used to assess reflection. An adapted version of Applebee, Langer and Mullis's (1990) scale with four levels was used to assess text elaboration.

In the first part of their study, the *portfolios were read holistically, one portfolio at a time.* "A total of 1,053 texts from 329 portfolios were read and scored. An average of 3.2 texts was read from each portfolio; an average of 27.4 min was spent per portfolio. At the end of the week, a 10% random sample of portfolios was read a second time to check reliability". (p.59) The first approach produced low levels of reliability. Agreement levels ranged between 25% and 58%, according to which scale was used. The researchers concluded that the initial training for raters had been adequate "but that raters had drifted apart over the course of the week. Further, we discovered that readers had had difficulty sustaining consistency in the face of continuously shifting topics and genres, not only among the different classes but also within the same portfolio" (p.63).

In the second part of the study, *texts were read by task across portfolios*. Moreover raters worked on all the portfolios from one course before moving to those of another. "Raters also spent time calibrating with one another by comparing their ratings on a few texts before proceeding" (p.64). The scales for scoring were also clarified. "In all, 939 texts written by 313 students … were read and evaluated. Each reader spent approximately 4 hr in training and 15 hr in reading and evaluating portfolios. Each portfolio required approximately 17 min to read…" (p.66). The mind boggles at the speed at which raters managed to assess many complex texts over various tasks. Interrater agreement was far higher than in the first study. Agreement levels ranged between 68% and 61% according to which scale was used. The researchers concluded that the relatively minor changes to the methods of scoring produced this far greater interrater agreement.

The researchers further concluded that it was damaging to reliability to ask for a holistic writing score for the portfolio in that writing quality varied so much according to the type of writing required. Readers of this research might also wish to conclude that it is not surprising to find significant differences in grading when assessors are asked to grade complex tasks at speed from a range of subject areas.

Nystrand, Cohen and Dowling (1993) in their assessment of college writing made the point that students may be penalised if the type of writing required in a portfolio are the types at which they are not skilled. Conversely other students may be advantaged if they are asked to produce writing which matches their strong points in writing. One could presumably argue, as hermeneutic researchers do that in this case, students must be allowed to select which of their texts will be assessed and that in a conventional examination, students will be even more disadvantaged as they will have less choice about what they are to be assessed on than potentially in a portfolio. At most, they will be allowed to choose between a few restricted questions in a timed examination.

Huot, Brian. "Towards a New Theory of Writing Assessment." *College Composition and Communication* 47 (1996): 549-566.

DeRemer, Mary. "Writing Assessment: Raters' Elaboration of the Rating Task." *Assessing Writing* 5 (1998): 7-29.

There has been a little research on how people actually use scoring rubrics. It is usually implicitly assumed that raters use scoring rubrics in a rational, consistent way. However, some researchers, mainly in American higher education writing programmes, have asked raters to talk aloud as they rate scripts (e.g. Huot 1993; Pula and Huot 1993; Vaughan 1991; Wolfe and Feltovitz 1994; Wolfe and Ranney 1996). The transcripts of the talk-alouds are then analysed. Usually the assessment task is relatively short, and involves grading

single pieces of writing (e.g. DeRemer 1998). To do anything longer would probably be intolerably intrusive for the rater, although very useful for a researcher.

In discussing research he and Judith Paulo did on placement assessment (Huot 1993, Pula and Huot 1993), Huot (1996) described how "raters tended to report making placement decisions not upon the established scoring guidelines on a numerical rubric but rather on the 'teachability' of students". He argued that "the context for reading student writing appears to guide raters regardless of rubrics or training found in many assessment practices" (p.554). Smith reported that "… talk-aloud protocols of raters using holistic methods for placement demonstrate that often raters first decide on student placement into a class and then locate the appropriate numerical score that reflects their decision (Smith cited in Huot, 1996 p.554).

At the risk of descending into anecdote, but after reflection stimulated by reading about the above research, I can present the following vignettes to describe my use of a scoring rubric used for both summative and placement assessment in a writing programme in an American university in the Middle East.

Learning to use the rubric – a training session

I sat on my wooden chair with side arm, together in a room with seven other new programme recruits and an experienced rater, our trainer. Rather nervously, I clutched my essay samples and my copy of the scoring rubric, one page of A4. Across the page were six or so scales (e.g. organisation, content, mechanical features). Each scale had a different possible total. Each one had pieces of text underneath describing what text features would merit what mark. At the end, one totalled these different elements. Maximum score 100. Minimum possible score 33 (or something like that) for reasons unclear to me.

Ready, steady, go. We set off to grade the essays. Mmh, when I had got over the initial stage of words in the essays dancing meaninglessly before my eyes, the essay seemed quite good to me. Much better than the secondary school essays I had been used to in Oman. There I thought myself fortunate if I got a paragraph that read coherently. I looked at the rubric. Yes, well organised, clear development of ideas – let's give it a – 12/15. Content – mmh – I scrutinise the rubric - ideas seem quite interesting, quite good – 15/20. … Total 85. I do another – 79. And another – that's really good – 92. Oh dear I seem to have run out of time. Others have finished the whole batch. I seem to be rather slow.

We all have to say the grades we have given for the essays. Oh dear, mine seem rather high. These people have high standards. People look at me. This is a little intimidating, not to say embarrassing. The trainer says sympathetically but firmly that I will learn. One of the other new recruits has problem grades as well. I feel a bit better. We discuss the essays individually. I get the message. Scores over 75 won't do. I'd better be more severe next time round. "Good development of ideas" in the rubric clearly means something different for these folks from what it means for me.

OK this batch has a nice range of marks – 49 to one 74. Oh no. Apparently I am now too low. It should have been about sixty. Everyone looks at me. The other rogue grader is in line now. We all discuss each essay.

Suddenly light dawns. I get the idea – the marks have to be between 60 and 75. Why did someone not say?! This is the range where these students "should" be. It isn't really between 33 and 100. Cunningly I work it out. If I give a mid-range mark, this public embarrassment will stop. If I ignore the rubric (which doesn't seem to relate to textual features I can recognise anyway), I can go faster. The important thing is to get a final score that is within the accepted range and to go back and alter individual elements of the overall score if the total is out of the range.

Tomorrow apparently I am to be let loose on grading two batches of placement tests, one grader among 30 others. I hope I can keep up with the speed of the others. I hope my grading partner feels a bit more competent than I do about this. Poor students. (BJ 1990)

We note here the pressures to conform acting on graders, the need to work at speed, the different meaning of the rubric for different graders and the essential disregard for the textual content of the rubric.

Further articles on reliability issues, not reviewed here individually are: Herman, Gearhart and Baker 1993 [writing, schools US]; Koretz (1998 [mathematics and writing, schools, US]; Koretz, Stecher, Klein, and McCaffrey 1994 [writing and mathematics, schools, US]; Pitts, Coles and Thomas 1999 [general practice trainers, medicine, UK]; Pitts, Coles and Thomas 2001; Underwood and Murphy 1998 [language arts, schools, US]; Valencia and Au 1997 [literacy, schools, US];

Suggestions for reducing interrater variability

The literature argues that certain actions can bring down the rates of interrater disagreement. Suggestions usually focus on matters such as:

- defining portfolio tasks more clearly and precisely (Davis et al. 2001);
- better training of assessors (Baume and Yorke 2002; Centra 1994; Davis et al. 2001; Nystrand, Cohen and Dowling 1993);

- the creation of a clear scoring rubric/descriptive statement *and/or* exemplars of expected outcomes (e.g. Davis et al 2001; Polyani 1962, Sadler 1989, Wolf 1995 all cited in Baume and Yorke 2002; Nystrand, Cohen and Dowling 1993);
- disaggregating the contents of portfolios so that tasks may be scored independently of other tasks (e.g. Nystrand, Cohen and Dowling 1993; Reckase 1995). Baume and Yorke point out that disaggregation may not be possible in all cases (e.g. with the Open University portfolios they describe or with ILT portfolios), depending on the organisational scheme for the assessment and so on. Hermeneutic assessors, with their concern for holism and context would argue that disaggregation is undesirable.

All these moves are intended to increase standardisation. Another suggestion has been to employ "fuzzy logic" in calculating portfolio grades (Fourali 1997). This is a quantitative technique which allows a rater to specify a range of grades they would find acceptable. If there is more than one rater, this provides a potential overlap in scores.

Some people argue that in view of the "low reliability" of portfolio scores, it might be advisable just to consider them for formative assessment (e.g. Reckase 1995).

In many cases where low agreement between raters has been reported, this is hardly surprising in that often graders have had little training, and they have been confronted with difficult and often decontextualised writing tasks. In cases of higher interrater agreement, such as in Baume and York (2002) and Nystrand, Cohen and Dowling (1993), training in grading tends to have been better and the grading rubric more carefully thought out.

Gaps in the research on interrater reliability in summative assessment in the positivist tradition

It is worth noting that there is little follow up research to see if these recommendations for better training and clearer rubrics actually work. Would raters actually "improve" if given better guidance? At least in higher education contexts, researchers tend not to do longitudinal research in this area. Nystrand, Cohen and Dowling (1993) is a rare example where a first grading study was followed up by another and procedures amended to see how this affected grading outcomes.

More research using "thick description" (Geertz 1973) would be helpful for investigating how raters actually use scoring rubrics. Avenues likely to be fruitful for investigating rater use of grading rubrics would be retrospective and/or stimulated recall interviews as well as other forms of detailed reporting of how the rubrics are used. See Smagorinsky (1994) for discussion of the use of such investigative tools. Such research would illuminate the murky areas of how raters actually use rubrics and how this relates to how assessment theory says that they should be used.

Some of the research is carefully done and well worked out, producing useful insights (e.g. Baume and Yorke 2002; Supovitz, MacGowan and Slattery 1997). However, most of the researchers in this positivist approach seem unaware of the theoretical and philosophical framework in which their work is rooted. And some of the research is carried out somewhat mechanically. For example, Centra (1994) in his study of the reliability of portfolio assessment for deciding on faculty contact renewal and promotion, discusses the procedures used in the assessment and in his research to establish what interrater reliability was. However, the research does not discuss inter-faculty politics and affective impact on faculty which presumably must have been an integral and significant part of this assessment process. The research does not discuss the validity of this method of assessment, a crucial issue when such high-stakes summative assessment depends on the assessment process. These are major omissions if the reader is trying to decide whether this would be a suitable assessment to use.

One issue to bear in mind when considering the statistics for interrater reliability is what are the likely levels of reliability for different rates on traditional UK essay examinations where essays are written on different topics, often in bad handwriting and marked by raters without a supporting rubric and often with inadequate training.

We move now to a discussion of research on agreement over rating in interpretivist approaches to assessment.

Agreement in interpretivist assessment of portfolios

Interpretivist researchers do not see disagreement and divergence of views in the assessment of writing as anything other than entirely normal. They do stress the notion of "interpretive communities which exist, or must be developed, in order to achieve intelligent meaningful assessments of portfolios. Discussion of values and standards, discussion of individual assessments of a portfolio and their modification in the light of discussion, and transparency of process are characteristics of this approach.

Reviews of specific studies

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

Broad (2000) carried out an ethnographic-type study of assessment in a Freshman Writing programme in a university in the United States. The purpose of the programme was to certify that students were proficient in reading, writing and critical thinking. It was a large-scale programme with 50 instructors. Students had to produce five essays, four of which went in the portfolio at the end of semester. The teachers started off wanting a way to grade the portfolios which was fast, easy and standardised. The process turned out to be rather more complex than they had hoped. Broad describes the assessment process, as it would be viewed through a psychometrically-based assessment lens and then through a hermeneutic lens.

The instructors and administrators attempted to "norm" essays (i.e. agree on assessment decisions for sample texts and portfolios). In large group "norming" sessions instructors had to give pass or fail grades to the texts and portfolios and discuss their reasons for doing so. Following on from this they met in "trios": one member being the class instructor and the other two members instructors of other classes on the same programme. In the trios, the "normed" instructors made the decisions whether to pass or fail the borderline portfolios from their group. This process of standardisation proved problematic in that firstly instructors could not agree on standard sample texts, finding them all individual and different. Moreover, different instructors valued different attributes in a text. In one trio one instructor felt strongly against clichés, while another insisted on correct technical conventions and another was more concerned about the vividness of the writing. Interpretations shifted around as the discussions continued. There was an essential conflict between the instructors wishing to be "reliable" and "standard" in their judgements while at the same time honouring diversity and complexity in writing. At times, Broad notes that the process did work smoothly.

Broad suggests that this diversity of judgement and the conflicts involved in reaching decisions would be viewed as undesirable from a psychometric viewpoint, but that from a hermeneutic viewpoint this was all very normal and part of the complex process that has to be gone through to reach sound decisions about writing. He suggests that in order to maintain "a sense of coherence in the program" without "insisting on a single standard" the programme should:

- rename its "norming" sessions to "articulation" sessions where instructors and administrators should speak out and explore both what unites and what divides them.
- allow instructors instead of administrators greater say in the articulation sessions in order to allow for other powerful voices to speak in the process of articulation.
- instead of choosing portfolios which are examples of evaluative decisions, choose instead examples of particular evaluative issues or problems and to expect that different people will have different views.
- publicise and report on evaluative pass/fail decisions in a brief manageable format so that everyone including instructors and students would understand the basis of the decision.

Allen, Michael. "Valuing Differences: Portnet's First Year." *Assessing Writing* 2 (1995): 67-89.

Allen (1995) discusses a fascinating project where he asked people from a variety of universities and geographical locations in the United States to embark on an assessment of the portfolios in one another's programmes. This did not involve high-stakes summative assessment – that had already been done for these portfolios – but was rather in the nature of an experiment arising out of a bright idea. Allen asked the volunteer participants to select 5 to 10 portfolios from their own department, to write a description of the programme in which these portfolios were produced and to send these to a central point for distribution. Each portfolio set and programme would be commented on by two outside readers. As Allen says, his article "describes the establishment of a hermeneutic circle of portfolio researchers acting as 'outside readers' for each other" (p.68). Allen went ahead with the project despite some discouraging comments about the difficulties in obtaining agreement on scores and comments.

Allen found differing rates of agreement at different stages of the project and for different types of portfolios. In the case of the first round of assessment of portfolios produced by programmes which were able to pass on rubrics and programme descriptions, assessors were able to reach an agreement rate of 82.5%. This high rate of agreement was probably encouraged as most of the portfolios had to be graded pass/fail, although as Allen points out several participants chose to send out portfolios for assessment which had been problematic for the local assessors. On the portfolios only operating in individual classrooms, there were far lower rates of agreement. Allen attributed this low rate to a complicated grading formula from one of the participating institutions which even he could not understand and to the fact that the students had been writing for their classroom teacher in a context those in the classroom understood, but which was mysterious to those outside.

In the second round of the project, conducted by e-mail discussions, Allen reported extensive discussions in which "most readers came to a rough evaluation and score before the session, but were willing to examine their readings and their assumptions in a polylogue of viewpoints, interpretations,

questions, and speculations about rhetorical intentions and educational priorities" (p.79). In this round, greater agreement emerged on the classroom portfolios assessed. Allen describes the nature of the discussion on the mailing group:

> E-mail tended to "draw out" our thinking, to encourage us to explore and develop ideas in ways unconstrained by convention or location. … There wasn't the same need to "calibrate" our readings so that we all read the same way, as there often is in a more formal "training session"; moreover our scoring was as much a discussion of meta-issues (how should we read a reflective letter?) as it was about scoring the portfolio, and certainly the "talk" was far more varied (informal, formal, analytical) than most training or scoring sessions we had been involved with (p.81-82)

Allen suggested several potential explanations for the high rates of agreement in this project where he had expected extensive disagreement. Examples of these potential explanations are: (1) readers are so used to scoring rubric guidelines that even if they do not agree with them they know how to tune in and use them; (2) e-mail discussion provided an environment whereby views could be thrashed out in a non-threatening and ultimately productive way; (3) traditional angst about disagreement would be far less necessary if psychometricians left teachers alone to reach agreement instead of forcing assessment structures on teachers with which they disagree. Allen cites a comment by one participant Marcia Dickson:

> I'd like to propose that the "statistics" on assessment are wrong. I know that we're not supposed to be in such agreement. However, whenever this group and other groups I've worked with have dealt with portfolio assessment, they have nearly always agreed on a grade/proficiency score/placement but not on the same merits (or demerits) of a portfolio. In other words - they agreed on a final judgement but for different reasons. Therefore: maybe the disagreement that is supposed to happen is a factor of trying to train teachers to evaluate by artificial and strict "standards" (a rubric done in the holistic ETS manner rather than letting them discuss and come to their own conclusions). The assessment technique leads to the disagreement, not the diversity of teachers' values (p.81).

Allen highlights the importance of context and expert reading in his explanations and asks an important question:

> One [broad idea] is the importance of context to evaluation, and the relationship between context and "expert reading". Does context, in the form of a well-articulated rubric, make for expert readers, or are expert readers needed in order for the local context to be understood (p.82).

Among the outcomes of the project, Allen describes how those who participated became more "self-reflective". They also became more aware of the portfolio contexts of other participants. Some portfolios were informal classroom portfolios while others were based around clear guidelines of the characteristics of each piece of writing, in effect "a collection of individual criterion-referenced tests" (p.76). Interestingly and perhaps tellingly, the two graders of this particular type of portfolio disagreed fundamentally with the principles underlying this type of portfolio assessment, but were nonetheless able to reach high levels of agreement on the portfolio ratings.

Allen discusses the implications of this project for portfolio assessment.

> Although persons more steeped in the professional language of assessment may find my use of the terms naïve, a procedure could be established for a socially constructed, agreed-upon way of establishing "fairness" in evaluation, if not classical "reliability." A sample set of portfolios could be sent by a local portfolio programme to outside readers who would read the contextual information (program description, rubrics, sample scored portfolios) and score the portfolios, and the outside readers and local readers could "talk" over E-mail about their scores, perspectives, differences, and suggestions. This "reliability" from the ground up, constructed out of local priorities and needs but also balanced by outside perspectives and suggestions. Indeed, it may be that "reliability" should be understood not as a function of scoring agreement, but rather as a function of the richness in reading and interpretation (p.83)

Allen suggests new terms to describe assessment processes. He talks about "shared evaluation".

> As a term "shared evaluation" implies social construction even as "validity" and "reliability" imply the ideal text against which individual student texts are evaluated. As we have practiced it (and *if* we have practiced it), a shared evaluation rests upon the following features:
>
> 1. A tentative score based on an evaluator's reading of a student text.
> 2. An openness of evaluators to other evaluators' perspectives.
> 3. An exchange of discursive analysis of the student text.
> 4. An examination of assessment issues that may arise from the exchange of evaluators' perspectives (p.84).

A further article on agreement over portfolios scoring from an interpretivist perspective, not reviewed individually, is Delandshere and Petrosky (1994).

Gaps in the research on agreement over summative assessment in the interpretivist tradition

The research literature on portfolio summative assessment from an interpretivist point of view is not nearly as extensive as the literature on inter-rater reliability from a positivist, psychometric point of view. The field

where such research has been done tends to be in the field of writing in the United States. There may be other such research, but I am not aware of it.

The implications for interpretivist assessment is particularly lacking in science areas. It would be useful to explore how interpretivist assessment might be applied to science. In other subjects, which clearly involve judgements about the quality and technical skill involved in creating the assessment artefact, it is relatively easy to see how interpretivist assessment might be applied, although the actual research probably does not yet exist. In science subjects, aspects of assignments which involve judgement (e.g. creativity, problem-solving and project work) would presumably be amenable to an interpretivist approach.

Some conclusions about agreement on portfolio assessment

In some ways, researchers working in the different assessment traditions are locked into fundamentally different ways of thinking and researching assessment. However, in practice as is so often the case they often appear to be moving closer together than the theoretical positions would indicate.

Researchers working in both positivist, psychometrically-based and hermeneutic traditions seem to be saying that with certain conditions in place, "good enough" agreement on portfolio assessment is possible. Those working in a psychometric tradition tend to stress the importance of clear grading rubrics and good training of assessors (e.g. Baume and Yorke 2002; Centra 1994; Davis et al. 2001; Herman and Winter 1993; Nystrand, Cohen and Dowling 1993; Supovitz, MacGowan and Slattery 1997). On the other hand, those working in a hermeneutic, interpretive tradition tend to stress the importance of inter-rater discussions, contextual knowledge and the holistic assessment of the entire portfolio rather than individual elements (e.g. Allen 1995; Broad 2000; Huot 1996).

In practice, those working in a psychometrically-based tradition often mention factors which show concern for social and educational contextual factors in assessment (e.g. Baume and York 2002; LeMahieu, Gitomer and Eresh 1995; Supovitz, MacGowan and Slattery 1997). Moreover, far from demanding exact agreement in scoring, and by accepting interrater correlations of around .8, those working in this tradition are effectively accepting that "ideal" scores, based on a reality external to the assessors, do not exist. Interpretivist assessors tend to argue that those in the positivist, psychometrically-based tradition use scoring rubrics and systems imposed from above. However, we can see in the case of Baume and York (2002) that although course designers give out a grading scheme to the tutor assessors, this is done along with a process whereby tutors are monitored in the feedback they give to try to ensure that feedback follows similar lines and whereby assessors attend meetings to discuss scoring guidelines, exemplars and problems that may arise. To take another example of positivist, psychometric assessment moving closer to the interpretivist position, LeMahieu, Gitomer and Eresh (1995) stress the need for the creation of an interpretive community.

> What we … consider central … is the shared understanding of the rubric by the raters. The rubric is ultimately a set of potentially ambiguous words on a page. Its meaning inheres in how people understand and apply these words. To the extent that such understanding is widely shared, there is a greater chance than that raters will apply the rubric consistently. They key feature of any performance system is this shared understanding or interpretive framework that members of an educational community bring to the assessment task (Gitomer 1993) (LeMahieu, Gitomer and Eresh 1995, p.25).

LeMahieu, Gitomer and Eresh show awareness of hermeneutic arguments, citing Moss 1994, and in fact, blend ideas about the need to develop an interpretive community when scoring portfolios. In fact, they present a holistic conception of the need to have integration of curriculum and assessment at various levels of the educational system. They include teaching practice, professional development, curricular development as well as assessment in their framework.

> This analysis suggests that, contrary to some findings and interpretations of previous efforts, portfolio assessment can have sufficient psychometric integrity to support purposes of public accounting. That said, it is clear that attaining this quality of information requires much more than specifications for structuring portfolios and for sound scoring practices. Four primary and related, components must be in place. First the purposes of the assessment must be clear, and the practices consistent with that goal. For example, if one is trying to understand the performance of the system, then efforts to get a number for every students are likely to have a negative impact on valuable opportunities to learn. Second, there must be a shared interpretive framework within the community conducting and using the assessment. This shared understanding is achieved through a hermeneutic approach and, when carried out in a disciplined way, can satisfy psychometric concerns as well. Third, there must be coherence in the system, so that accountability goals are consistent with classroom goals. This is not to say that the type of information available at various levels of the system will not differ – just that the basic values implied by varying purposes should not be at odds. Fourth, and finally, the development of this portfolio-based instructional and assessment system was dominated throughout by conversations designed to address issues of quality in both instructional and psychometric terms.

> Whereas traditionally the conversation typically excluded one or the other viewpoint, in this case both were pursued with equal diligence (LeMahieu, Gitomer and Eresh 1995, p.28).

On the other hand, a hermeneutic researcher such as Allen (1995) specifically argues that "shared evaluation" based on interpretive, hermeneutic principles "cannot substitute for large-scale assessment programmes, but it could be that an outside reading, response, or check (the choice probably depending on one's philosophical orientation) can lead a local assessment procedure to increased fairness" (p.84).

Psychometrically oriented researchers tend to stress the importance of assessing tasks individually. For example, Nystrand, Cohen and Dowling (1993) advocate the grading of single assignments at a time across portfolios in order to achieve high rates of agreement. However, other researchers working in this tradition such as Baume and York describe a system where assessment is carried out across the portfolio according to different "outcomes" and "underpinning values".

Some researchers working in the positivist, psychometrically-based tradition address the issue of fairness in a tradition which demands that the same be required of each assessee. Baume and York (2002) address this by pointing out that it was harder for some candidates than others to produce evidence of work towards equal opportunities in their portfolios. Nystrand, Cohen and Dowling (1993) make similar points. These are issues that would equally worry hermeneutic assessors.

Other issues of interest arising out of these particular articles are as follows. Articles describing the assessment processes where portfolio assessors discuss their judgements with others frequently talk about the professional growth such discussion entails (e.g. Allen 1995; Supovitz, MacGowan and Slattery 1997).

Neither research tradition includes much action research. Assessment would be a very suitable field for such enquiry.

Agreement between raters is just one aspect of portfolio assessment. Claims about learning through formative assessment is another important aspect which we will now discuss.

Section Three The links between learning, formative assessment and portfolios

Claims, issues and questions

Many claims are made about the opportunities that portfolios present for learning through various forms of formative assessment. How far does research confirm that these opportunities exist? What is the nature of these opportunities? What problems exist? What gaps are there in the research? In examining this claim and answering these questions, it is important to unpick various issues and claims.

The major claim with which we will be concerned is that **formative assessment of portfolios can enable productive forms of learning to take place**. The forms of formative assessment considered in this section are:

- self-assessment in the shape of reflection
- peer assessment
- teacher-led feedback or dialogue with the student.

Please note that the term "feedback" is rather unsatisfactory as it implies a one-way process of information flowing from teacher to student, where the student passively receives the information, so "dialogue" between student and teacher may be preferable, although this does not convey the notion of the teacherly input and expertise which inevitably form part of a formative assessment process.

Sub-issues to probe when considering the major claim about the learning opportunities offered by formative assessment of portfolio are:

- is it possible for students to reflect (self-assess) to such an extent that they actually learn in ways that are productive?
- does peer assessment of portfolio content actually move a student's learning forward?
- does teacher-led formative assessment of portfolio assessment move student's learning forward?

A further issue to consider is whether it is *possible* for there to be summative assessment of learning as evidenced in a portfolio. Such an assessment is likely to be complex and highly subjective. How does this sit with conclusions drawn in discussions in Section Two of this report on interrater agreement? It is important to note that the two bodies of research on (1) *interrater agreement* (discussed in the previous section) and

(2) *summative assessment of learning having taken place* (discussed in this section), as evidenced in portfolios, operate quite independently. Most of the articles where summative assessment of learning development is a concern are not interested in issues of interrater agreement. They focus instead on whether learning has or has not taken place. Some articles take a broader approach and discuss both issues of learning and interrater agreement. However, these tend to treat both subjects somewhat more superficially. The article by Davis et al. (2001), reviewed later, is of this latter type. We have included it as it seems an interesting article for those seeking to introduce portfolios and it seems a typical format for articles on portfolio assessment in medicine. It covers a wide range of problems and solutions, but does not focus in depth on one area.

A further area to investigate is the claim made about portfolios that they permit "authentic assessment" which should have greater predictive value about how students will actually perform in professionally related areas, for example. Is there any evidence of this?

Questions to bear in mind when answering these questions are:

A. is it formative assessment of work that happens to go into the portfolio that allows the learning to take place?[12] In this case, the portfolio itself is a *passive receptacle* which is probably produced for summative assessment at the end. It seem legitimate for us to discuss this type of assessment situation in that the portfolio with its range of work produced for consideration in summative assessment allows, although it does not necessitate, formative assessment of various kinds to take place. Other methods of assessment (e.g. timed examinations) might not allow such formative assessment to take place.
B. Or is it formative assessment, most probably in the shape of reflection on the part of the student him/herself and perhaps enhanced by dialogue with the teacher, about the *portfolio contents* and *construction* that enables the learning to take place? In this case, the portfolio itself is *an active vehicle of learning.*

Some background: learning theories and assessment theories

In order to probe these questions, it will be useful to examine some background theory to the issues involved. Learning, teaching and assessment are intimately bound together in these discussions. As Murphy (1994) describes:

> … with a constructivist framework, students are seen as active collaborators in the building of knowledge. Learning takes place through interaction, existing in the transaction between student and student, student and text, student and teacher. Viewed from a constructivist perspective, then, assessment procedures are inevitably a part of the dialectic of teaching and learning, part of the process which defines what knowledge is, what is learned, and how students learn. Assessments which reflect this perspective provide a means for engaging students in self-reflection and for acknowledging their role as collaborators in the learning process. In sum, a constructivist perspective acknowledges the reciprocity and interdependency of assessment and curriculum (190).

The discussion of learning theory below is not exhaustive, but rather illustrative of relevant work. In this section of the discussion, I will draw heavily on the notions of Vygotsky and experiential theorists. The aim in the discussion is to show how theories of learning might relate to one or more of the types of formative assessment mentioned above.

Readers who wish to pursue issues of learning, especially as they relate to assessment, in more depth, may wish to consult Brown, Bull and Pendlebury (1997) and Heywood (2000).

Vygotsky and learning

Vygotsky was a child psychologist whose ideas have been adapted for adult learning. He died more that 60 years ago, and his work has been continued by neo-Vygotskians such as Jerome Bruner (e.g. 1985) and David Wood (1988), as well as many Russian researchers. They have developed such notions as 'scaffolding', staged adult or expert peer support for children's learning (cited in Mercer 1996, p.30),

The indirect relationship between teaching and learning

In this view, the relationship between teaching and learning is not direct. On the contrary, it is very complex. Learning is an active and a social process. The teacher strongly influences a classroom environment and orchestrates suitable presentations, continuing graded practice activities and opportunities for various types of interaction within the classroom. However, these are just initial staging posts, pre-conditions to learning.

[12] See previous discussion of this issue in the section *What is a portfolio? What are the purposes of portfolios* p.37-39.

The student has to do the actual learning and internalising. After initial teaching, students may be able to do the activity or they may forget it, or not have absorbed it or have partially absorbed it. As Vygotsky (1962) explained:

> Instruction has its own sequences and organisation, it follows a curriculum and a timetable, and its rules cannot be expected to coincide with the inner laws of the developmental processes it calls to life. On the basis of our studies, we tried to plot curves of the progress of instruction and of the participating psychological functions; far from coinciding, these curves showed an exceedingly complex relationship (p.101).

Vygotsky posited a sequential relationship between instruction and development as follows:

> When the child [adult student] learns some arithmetical operation or some scientific concept, the development of that operation or concept has only just begun. Our study shows that the curve of development does not coincide with school instruction; by and large, instruction precedes development (p.102).

The contribution of teachers or more capable peers to learning

Although Vygotsky does not perceive teaching as directly effective, teachers do have a central role. For Vygotsky (1962), cognitive capacity unfolds with instruction. Without instruction, these abilities will not just naturally develop as they are culturally and socially shaped and have to be learned. The role of the teacher, according to Vygotsky is to provide a structured learning environment and he offers valuable insights into how minds develop and how to structure the environment to promote learning. Vygotsky (1978) talked about the zone of proximal development which he defined as:

> the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers (p.86).

He argued that the learner, through structured learning activities *undertaken in interaction with* the teacher or more capable peers, would progress through the zone of proximal development which would become the level of actual development, whereupon the learner could then proceed through his/her new zone of proximal development. As Vygotsky (1962) wrote, “What the child [adult student] can do in co-operation today he can do alone tomorrow” (104). This educational scaffolding involves structuring tasks through instruction, modelling, questioning, and feedback until the learner can operate independently (McCarthey and Raphael 1992, p.17-18).

The active nature of the learning process

Through exposure to a learning environment, the student makes meaning of information by relating it to past and present experience, past and present information and future expectations in an active dialogue which involves internal dialogues within each student and dialogue within and outside the classroom between teacher and student, teacher and students, and probably with numerous other individuals. As Vygotsky describes of those who have reached a conceptual stage of development, learning “proceeds by the interaction of generalising propensities of conceptual language and empirical experience” (cited in Cope and Kalantzis 1993, p.70). “Conceptual knowledge proceeds, in other words, via the dialectical interplay of induction and deduction, and of theory and experience” (70). This process will be individualised as each student will have been exposed to different experiences and different information in the past and will react differently in the present context. In this dynamic, dialogic conception of knowledge, thinking is a dynamic social activity, gradually internalised by the individual (Vygotsky 1978).

> Vygotsky (1978) argues that thinking is a social activity, initially shared between people but gradually internalised in the individual. Cultural artefacts (language in particular) are experienced through interaction, re-enacted by the individual, and eventually internalised. In a similar vein, Brown et al. (1989) call learning a matter of *enculturation*: by observing and living within a particular culture, people gradually start to adopt the behaviour and belief systems of the culture (Perkins, Jay and Tishman 1993, p.16-17).

While students are in the process of learning, one can expect them to participate in the learning dialogue only in a very tentative questioning manner initially until they have built up a sufficient knowledge base to participate more fully. However, during this early period, according to a Vygotskian analysis, the teacher has to continue to give the students structured activities to enable them to practise participation in order to facilitate their gradual building up of appropriate skills. Formative assessment is likely to be very useful in this process.

Experiential learning theory

Learning and re-learning

The work of experiential theorists, influential in adult learning, is also helpful in understanding the nature of learning by suggesting reflective learning cycles. Experiential learning theorists posit that individual experience, information gathering, and reflection thereupon drive learning and action forward and upwards in a learning spiral (Kolb 1993). Tensions between immediate personal experience and abstract concepts alter personal concepts (Kolb, p.146). Immediate personal experience is used to test and validate abstract concepts (Lewin cited in Kolb, p.140). As Kolb explains: "learning is the process whereby knowledge is created through the transformation of experience" (p.155). Learning is a continuous process:

> The principle of continuity means that every experience both takes up something from those who have gone before and modifies in some way the quality of those which come after ... (Dewey 1938, p.35 qtd in Kolb, p. 145).

Kolb points out the significance of this. "The fact that learning is a continuous process grounded in experience has important educational implications. Put simply, it implies that all learning is *re-learning* (Kolb, p. 145). The students all have their own theories and ideas already - they are not blank pages - and our role as educators is to modify these old ideas which well may conflict with the new ones we are trying to teach (p.146). As Kolb argues,

> ... one's job as an educator is not only to implant new ideas but also to dispose of or modify old ones. In many cases, resistance to ideas stems from their conflict with old beliefs that are inconsistent with them. If the education process begins by bringing out the learners' beliefs and theories, examining and testing them, and then integrating the new more refined ideas into the person's belief systems, the learning process will be facilitated (p. 146).

Class discussion and opportunities for individual formative assessment will therefore have a useful role in facilitating this relearning process.

Learning outside the immediate classroom

For experiential theorists, learning is not just a cerebral matter. Learning is a holistic process of adaptation to the world. Experiential learning is concerned with the whole human being - with thinking, feeling, perceiving and behaving. "Learning is *the* major process of human adaptation" (Kolb, p.149). And it is not just confined to the classroom, but extends to all aspects of life and work (Kolb, p.149).

> In addition to knowing how we think and how we feel, we must also know when behaviour is governed by thought and when by feeling. In addition to addressing the nature of specialised human functions, experiential learning theory is also concerned with how these functions are integrated by the person into a holistic adaptive posture towards the world (p149).

The role of context in teaching, learning and assessment

Context is recognised increasingly as being of great importance in teaching and learning. As Darling-Hammond and Snyder (2000) write:

> … all teaching and all learning is shaped by the contexts in which they occur. These contexts are defined by the nature of the subject matter, the goals of instruction, the individual proclivities and understandings of learners and teachers, and the settings within which teaching and learning take place. Such variables as school organization, resources, materials, amount of time and how it is structured for learning, the duration and nature of relationships among students and teachers, community norms and values influence the processes and outcomes of teaching decisions. The extent to which context influences teaching- and determines what kinds of approaches to teaching will be effective – is a factor that is increasingly acknowledged in research on teaching, in teacher education, and in the assessment of teaching (p.524).

In turn, educators have been looking for means of assessment which take context into account. Portfolios are one of the methods of assessment that offer access to information about context:

> … try to capture important attributes of teaching and reasoning about teaching. [They] allow the application of theoretical principles to problems in specific contexts while appropriately complicating efforts to draw generalizations about practice. In doing so, they may also transform teachers' understandings of both theory and practice (Darling-Hammond and Snyder p.525).

Implications of theory for teaching, learning and assessment

Taken together these approaches and understandings suggest both a useful theoretical understanding of what happens during learning, and also indicate specific areas the instruction programme should address: a carefully structured learning environment, individualised as much as possible; the need for space to reflect; for students to express, discuss, test and modify their ideas and theories; and the opportunities for students

to develop an accepted and authoritative voice in interaction with other voices. We should examine the description in the reviews of articles which follow to see if they have incorporated such elements into their portfolio construction and related formative assessment process.

What we can draw from this discussion from the point of view of assessment is that:

- The path of student learning is likely to be highly complex and individualised. Learning will take place over a long period of time and not be linear or straightforward. Given the idiosyncratic and individual nature of learning one cannot expect that different students will react in the same way to the teaching input they are given. They come to the class from different learning contexts, with different personalities, and different motivations. Formative assessment, individualised as far as possible, is likely to be useful in moving this complex process of learning forward, although one can never expect a direct input-output correspondence. Formative assessment may operate in the form of expert help assisting students to move forward through their zone of proximal development (Vygotsky) or in the form of helping them reflect and transform their experience (Kolb).
- Learning products will vary considerably between individuals because of the individualised nature of the learning process.
- Summative assessment of the learning which has taken place is likely to depend on the assessor being able to access the complex learning process.
- Through formative assessment teachers are given the opportunity to get feedback on their teaching and to adjust their teaching accordingly

Portfolios do seem to offer opportunities to ~~achieve~~ address these issues. As Wolf (1993) described, "Unlike more traditional assessment methods, portfolios are conceived not just as an end product, but as an episode of learning (cited in Supovitz, MacGowan and Slattery 1997, p.239). We have to examine the articles that follow to see how they address these issues.

Portfolios and authentic assessment

In addition, portfolios seem to be well suited to the needs of authentic assessment. Darling-Hammond and Snyder (2000) suggest four aspects of "authentic" assessment which seem likely to be important for assessing teaching and enhancing candidates abilities to teach well. Although they are speaking about teacher training, the points they make about assessment are applicable to many other types of professional practice. Also, although they are speaking about authentic assessment in general, their remarks could certainly be applied to portfolios as a form of authentic assessment. They suggest that it is important that:

1. Assessments sample the actual knowledge, skills, and dispositions desired of teachers [professionals] as they are used in teaching and learning contexts, rather than relying on more remote proxies.
2. Assessments require the integration of multiple kinds of knowledge and skill as they are used in practice.
3. Multiple sources of evidence are collected over time and in diverse contexts.
4. Assessment evidence is evaluated by individuals with relevant expertise against criteria that matter for performance in the field.

Assessment offers the opportunity to learn and practice "the desired outcomes and for feedback and reflection" (p.527-8). Portfolio assessment seems to offer these opportunities.

Further links between portfolio assessment and learning theories

Murphy (1994) discusses the links between assessment theory and learning theory.

> Portfolio assessments which give students greater authority and responsibility for demonstrating their learning and accomplishments provide an assessment model which is compatible with a constructivist perspective (p.192)

Murphy makes extensive claims about learning and assessment in portfolios as follows. She argues that:

> when students are allowed to make their own selections, whether in relation to standards, other externally defined criteria, or their own goals for writing, they are encouraged to engage in the self-reflective process of reviewing and evaluating their writing and themselves as writers (p.190-191).
>
> It makes sense to give students a voice in the assessment process, so that they have some stake in it, a stake for the decisions they are empowered to make,, not just for the consequences of failure. I do not mean, of course, that students should be making these decisions about portfolios without any guidance from their teachers. But students need to be able to exercise judgement, because it is in exercising judgement that students learn how to assess a piece of writing, or a whole collection of writing. If the contents of the portfolio are specified too narrowly, that is, if the students are simply

> given collections of assignments to write and put eventually in their "portfolios", they will have little room to exercise judgement or to collaborate in the building of knowledge (p.191).
>
> If portfolio assessments are meant to be compatible with new views of learning which emphasize the student's active role in constructing meaning, then students like teachers, will need to become full participants in the assessment process. From the student's perspective, portfolios can allow students to gain some control over the process, to demonstrate more completely in their own terms what they know and can do, and to set their own goals and assess their progress toward them. But if the contents of portfolios and the processes for generating them are externally mandated and highly standardized, students will have little chance to exercise judgement or develop responsibility. If one of the goals of our educational system is to enhance the development of students' individual initiative, creativity, and responsibility, then students will need to be allowed some authority in deciding how their work will be represented. In turn, assessment systems will need to accommodate some diversity in the ways that standards can be met (p.201).

Can students rise to these challenges in practice? Do we necessarily want students to be able to? Should we be more modest in our portfolio ambitions? Should we have different priorities?

Too constrained a portfolio system may be unhelpful for learning. Indeed, Murphy, Bergamini and Rooney (1997) and Callahan (1997) describe the constraints that an externally imposed portfolio system placed on a school based portfolio system. However, assessors may find Murphy's proposals too radical or unsuitable for local conditions. To take one such example, when I was teaching first year undergraduates in Cairo, students who had recently emerged from an educational system which priorities rote learning, they would simply have been overwhelmed and unable to cope with being deeply reflective on their work. To take another example, in medical contexts, the priority may be more to get students minimally competent at certain activities before allowing them to practise as doctors, rather than encouraging them to become very reflective. Or it may not. Snadden and Thomas (1998) concluded when discussing the potential for medical portfolios, that:

> while the emphasis on grading, excellence and comparison between students and doctors remains in assessment and medicine, it is likely that portfolios will have a greater place as a learning tool than as a summative assessment took, but as a learning tool they may have considerable influence on a learner's performance in summative assessments. If, however, they can be grasped to develop innovative assessments they may have the potential to influence the type of learning in medical curricula to encourage the development of reflective practitioners (p.197).

Areas where research on the learning potential of portfolios has been carried out

Portfolios have been adopted because of their learning potential in various areas, but have been concentrated in specific areas such as teacher training, medicine and writing. Some examples of this work are listed below.

Training of school teachers and university lecturers

A considerable amount of research has been done on the learning potential of portfolios in the training of school teachers and university lecturers (e.g. Darling 2001; Frederiksen, Spiusic and Sherin 1998 [video portfolios]; Simon and Forgette-Giroux 2000; Taylor 1997; Wade and Yarborough 1996). Portfolios are seen as having particularly useful potential for "represent[ing] a holistic view of student growth and the development of professional qualities that are key for effective classroom teaching" (Darling 2001 p.117) and to be a way of moving away from technical and mechanistic assessment of trainees. As Darling writes:

> There have been significant efforts to transform teacher education in recent years, efforts that have moved programs beyond the ubiquitous talk of teaching skills, techniques and behaviors. It is widely accepted that learning to teach is as much a matter of cultivating certain dispositions and sensitivities as it is amassing pedagogical tools (Zeichner & Liston 1987; Reichert 1990; Zeichner & Tabachnik 1991; Fenstermacher 1992; Tom 1997). Understanding the reasons for the curricular decisions one makes, being able to justify pedagogical choices, knowing both limits and possibilities of various methodological approaches, and meeting the unique and dynamic needs of learners, all require a critical eye, an inquiring spirit, and careful judgement. Clearly, a teacher certification programme is only the beginning of one's education for teaching, but it does represent the foundation for future exploration. While we cannot instil a passion for ongoing discovery and learning (including self-discovery) we can nourish it. Portfolios may help us do just that (p.119).

In teacher training, portfolios have been advocated as:

- promoting reflective practice (e.g. Borko et al. 1997; Wade and Yarborough 1996)
- a means of initiating dialogue about teaching and learning (Loughran and Corrigan 1995)
- providing evidence of achievement in learning to teach (Loughran and Corrigan 1995)
- providing a vehicle for teacher learning and development (Athanases 1994)

As Adler 1994 describes, reflection has been considered variously to be:

- deliberation on practical teaching matters (Cruickshank 1987)
- thinking about problems in the midst of teaching (Schön 1983, 1987)
- questioning institutional goals and criteria (Zeichner 1981; Zeichner and Liston 1987) (cited in Wade and Yarborough 1996)

Wade and Yarborough point out that although portfolios will not substitute for practical experience in schools (in this case), portfolios can help foster these types of reflection by focussing student thinking on key issues.

Presumably these arguments could be extended into other type of professional and vocationally oriented training programmes such as social work and medicine. In all such courses a major problem is how to move from theoretical and intellectual arguments to practice. Practitioners have to be able to engage in systematic learning from the complexities of practice in context as well as from disciplinary theory. A major issue is how to assess all of this. Presumably also the arguments are applicable to other types of courses which are not vocational as such.

Initial medical training and continuing professional development

In medical schools and with general practitioners in continuing professional development, there have also been pressures from the General Medical Council to "avoid information overload, develop a learner-centred problem-oriented approach and improve communication skills" (Finlay, Maughan and Webster 1998). Experimenting with portfolios has been a part of this process (e.g. Challis 1999; Davis et al 2001; Finlay, Maughan and Webster 1998; Mathers et al. 1999; Snadden and Thomas 1998). In medicine, especially in undergraduate education, a major concern has tended to be that students acquire specific minimal skills and knowledge in order to practise safely so portfolios have tended not to be as radical or open in orientation as in teacher training.

Challis (1999) describes various examples in UK medical education where portfolios have been introduced as part of the assessment process: Year 5 Undergraduate Module in General Practice and Community Health Care at the University of Sheffield; pre-registration house officers; general practice vocational training in one region in Scotland; specialist registrars: examples from paediatrics and public health; general practice trainers in Wessex; supporting continuing professional development of general practitioners in Sheffield; electronic portfolios: the WISDOM project; learning opportunities for teams ('LOTUS') – supporting the educational needs of general practice staff; postgraduate certificate in medical education at the University of Newcastle upon Tyne. These descriptions give a useful picture of the potential of portfolios for enhancing learning, but unfortunately there does not seem to have been detailed research of the effectiveness of these projects, or at least it is not mentioned in this article.

Writing portfolios

e.g. Gearhart and Wolf 1997; Johnston 1998; Underwood 1998.

Engineering

In undergraduate engineering, Payne et al. (1993) describe similar moves towards portfolios. Payne at al. introduced portfolios in their course at Sheffield Hallam university in response to an invitation from the DTI and the Engineering Council for proposals for 'Integrated Engineering' courses.

Training of managers

Smith and Tillema (1998).

Figure 5 offers an overview of issues addressed in each of the studies. *Figure 6* offers a description of the investigative methods used by the researchers. *Figure 7* offers a description of the subject area discussed in each article.

Figure 5

Issue examined	Study
Reflection	Darling (2001); Johnston (1998); McLean and Bullard (2000); Wade and Yarborough (1996)
The learning process	Darling (2001);Davis et al. (2001); Darling-Hammond and Snyder (2000); Johnston (1998); McLean and Bullard (2000)
Other opportunities offered by portfolios (e.g. scaffolded learning, connecting thinking with performance)	Darling-Hammond and Snyder (2000); Finlay, Maughan and Webster (1998); McLean and Bullard (2000); Wade and Yarborough (1996).
Student motivation	Darling (2001); Snadden and Thomas (1998); Wade and Yarborough (1996);
Professional growth	Darling (2001); Darling-Hammond and Snyder (2000)
Conflict between summative and formative assessment in portfolios	Snadden and Thomas (1998)
Influence of context	McLean and Bullard (2000); Johnston (1998)

Figure 6 Method of investigation (apart from scrutiny of the portfolios themselves)

Method of investigation	Study
Students essays	Johnston (1998); Wade and Yarborough (1996)
Videos	Darling (2001)
Standard mark sheets	Finlay, Maughan and Webster (1998)
Questionnaires	Davis et al. (2001); Wade and Yarborough (1996)
Open-ended interviews	Darling 2001; Johnston (1998); Wade and Yarborough (1996)
Analysis of tutorial discussions	Johnston (1998)
Pre- and post-tutorial discussions	Johnston (1998)
Verbal reports	Davis et al. (2001)
Various students notes and drafts	Johnston (1998)
Observer documentation	Davis et al. (2001)

Figure 7 Formative assessment in portfolios: field discussed in the article

Educational sector	Specific area (if relevant) and country	Study
Schools	Teacher training	Darling (2001); Darling-Hammond and Snyder (2000); Wade and Yarborough (1996)
Higher education	University teacher training	McLean and Bullard (2000)
	Undergraduate medical education	Davis et al. (2001); Finlay, Maughan and Webster (1998)
	Continuing medical education	Snadden and Thomas (1998)
	Writing	Johnston (1998)

Learning through formative assessment: Reviews of specific articles

The claims that can be made about the stimuli to student learning, arising out of formative assessment, must be examined with some care. It is not enough (when evaluating the impact of portfolios) to demonstrate that learning has taken place during a course or that students who have completed portfolios have higher grades at the end of the course unless a piece of research can demonstrate that this improvement is at least partly due to formative assessment which has happened during process of producing the portfolio.

As ever one has to be aware that many of the articles on the opportunities that portfolios offer for learning through formative assessment are not research articles. This report will only review articles that report on empirical research or detailed evaluations of specific projects.

School teacher training portfolios

Darling, L. Farr. "Portfolio as Practice: The Narratives of Emerging Teachers." ***Teaching and Teacher Education*** **17 (2001): 107-121.**

Darling (2001) investigated the portfolio production of 31 teacher trainees in Canada. The students had completed two terms of coursework by the time they handed in the portfolios. In the portfolios, although the students were encouraged to be highly creative in terms of format, themes and presentation, they were required to include four components: (1) an introduction explaining the reasons for selection of the materials, (2) a philosophy of teaching, (3) an example of teaching practice such as a lesson plan, (4) an action plan for teaching which included curricular goals and plans for professional development. In discussions with the students, the students said they preferred grades (excellent, good, fair, unacceptable) rather than just pass/fail grades given the effort the portfolios required.

Darling carried out her investigation through video recordings where 12 volunteer students discussed their experiences with portfolios after handing in the portfolios (as if to future students) and in an open-ended interview later in the year, after an extended period of teaching practice in a school. Darling asked the students to talk about the value of the portfolio to their teaching.

As Darling describes:

> For eight of the [12] students interviewed, the portfolio assignment turned out to be a unique story of their experiences in "learning to teach". Their emerging professional identities were documented, sometimes powerfully, through combinations of words and images expressing understanding, anticipation, intellectual vigor, fear, confusion, disappointment, empathy, and … exhilaration (Darling 2001).

For the other four students, external goods or extrinsic motivation were more important in the construction of the portfolio. Darling (2001) makes a distinction between two kinds of goods, associated with social practices.

> *External goods* – these are closely related to extrinsic motivation. (e.g. "things such as rewards, prizes, grades, and recognition, goods that are bestowed on a practitioner subject to the judgement of other persons such as employers, judges, critics, teachers, and sometimes the general public … most people making these judgements are themselves 'inside' the practice".
>
> *Internal goods* – these are closely related to intrinsic motivation. (e.g. "the pride that comes from accomplishment, and the pleasure that comes from engaging in certain sorts of activities skilfully and successfully … appreciation of the inherent value of a complex human achievement such as the elegance of form in a mathematical theorem or a piece of music" (p.109).

In order to achieve internal goods one has to be prepared to submit one's work for criticism. This kind of process tends to require certain virtues such as "justice", "courage", "honesty", tenacity and willingness to learn from mistakes (p.110).

Darling certainly found that eight of the twelve students (who all achieved either "excellent" or "good" in their portfolios) found portfolios a useful way of documenting and exploring their professional growth.

On the other hand, in her discussions of these Type B portfolios, Darling points out that although the use of portfolios has a "sound … theoretical base, the potential of portfolios has not always been realized in practice" (p.118). She described how in her own study:

> In some instances, portfolios remained random collections of undeveloped thoughts and ideas … As well, the potential "internal goods" of the portfolio were never realized for some students. Their own sense of growth and accomplishment as emerging teachers did not find expression in a meaningfully told narrative. Several students never moved beyond viewing the construction of a portfolio as another "hoop to jump through". In order to make the portfolio experience as meaningful as possible for our students, we need to address some important concerns, several of which were expressed by the students themselves and several of which came through in their work. (p.118).

Anxieties mentioned by Darling's students included:

- (early and in some cases continuing) anxiety about the scope and nature and value of the task;
- lack of "models" that might guide early phases of construction;
- little academic preparation for a creative and personal piece;
- concern about the subjectivity of the evaluation (p.114, 118).

To give an example of anxiety:

> Authenticity and sincerity were key elements for Louise who "fretted and worried" about expressing herself in genuine ways, "as a real person here". When students heard words like "imagination and creativity" associated with the task, many feared they had no experience or aptitude for, as one student put it, "writing poetry or making art". For many students, deciding on an approach "that felt honest" in Louise's words, and a method of organization that "would work for me" as Jenny put it, were worried from the beginning (p.114).

It is a little worrying to see the emphasis put on creativity, the use of metaphors, the cohesiveness required in the portfolio in this study. One can imagine that producing a creative, cohesive artefact rich in metaphors out of messy, and nascent professional experience could be hard and intimidating. One could imagine that worry over format of the portfolio, rather than everyday and strategic grappling with professional development could be a major anxiety for some people, especially those not used to working in this mode.

Wade, Rahima, and Donald Yarbrough. "Portfolios: A Tool for Reflective Thinking in Teacher Education." *Teaching and Teacher Education* 12 (1996): 63-79.

Wade and Yarborough (1996) investigated the portfolios of 212 undergraduate elementary teacher trainees in the United States. They used students' essays, in-depth interviews and anonymous questionnaire surveys to investigate the issue. The essays were reflections on the portfolio experience. The interviews involved seven students who had had varying degrees of enthusiasm about the portfolio experience.

The portfolios "are part of a required social studies methods course to document their learning and growth in the Youth and Elderly in Service (YES) community service-learning project. Throughout YES, students work in parts with one senior citizen and one child from a single parent family for one hour a week throughout the semester. The intergenerational foursome typically play games, read books, write stories, go on outings, make craft projects, and share the differences and similarities in their life histories" (p.66). The portfolios included both items required by the instructors and items selected by the students. "The five required assignments are: a reflection on the student's experience with the elderly, a research paper examining a national issue affecting the elderly, an intergenerational lesson plan, and two letters to the instructor. The mid-semester letter focuses on describing what is included in the portfolio at that point in time and why the student made the choices she or he did. The final end-of-semester letter is a reflective essay on the student's experiences with both [the course] and the process of creating a portfolio" (p.66).

Wade and Yarborough reported on the many ways that the portfolios helped the students reflect. Students wrote in their essays that portfolio construction helped them reflect on relationships, the ageing process, community issues and communication. Students also said that they learned "patience, optimism, open-mindedness, self-efficacy, and gratitude through their reflections" (p.71). Some students also made links between the YES course and other interests they had such as literature, poetry and volunteer work (p.72).

One should perhaps notice that this reflection might have been at least partially stimulated by course content and might have occurred at least partially even without the portfolio element. Perhaps the important point to note, rather than making a rather unhelpful distinction between developments which happened as a consequence of course content and developments which happened as a consequence of portfolios development, is that course content, learning and assessment method can be integrated in this useful way.

Wade and Yarborough (1996) found that many students had struggled initially with creating a portfolio. In the questionnaire, 60% of students "agreed or strongly agreed that they were confused initially about how to create a portfolio" (p.69). In the essays, 23% wrote of the students about their difficulties in working with such an open-ended and unfamiliar assignment, although most later overcame these problems. Two of the students in the interviews mentioned this kind of pattern as well. Two others spoke about continuing confusion. It is probable that initial struggles with portfolio should be viewed positively, as indicative of a time or growth (p.77). Continuing struggle and confusion is likely to indicate a process that is not working.

In the survey, 37% of students claimed that the portfolio did not help them to reflect on their learning experience. The researchers found that the quality of instructor's feedback affected the capacity of the student to reflect. Clearly that such a large proportion not finding that portfolios helpful in the reflection process is a cause for concern and one that would warrant caution.

Wade and Yarborough found five factors which tended to contribute to student confusion and anxiety about portfolio construction:

- nature of the instructor's initial introduction of the portfolio and subsequent feedback
- lack of [previous] exposure to portfolios,
- the open-ended nature of the portfolio,
- expectations about education coursework – some students expected to work out what instructors "wanted" and to focus on achieving that – they may also have expected discrete assignments
- students not investing their energy in the construction process because of various reasons such as unhappiness with the course overall, a belief that a portfolio should be an option rather than a requirement (p.76).

They summarise their findings as follows:

> Portfolios are probably not the best tool for professors who start the semester with a prescribed list of A to Z facts they hope their students will gain from their courses. Rather, the portfolio is a potentially valuable method in teacher education programs that are based on the constructivist notion of students learning from experience, creating their own meaning, and developing both expertise and commitment to the process of reflection. While portfolios are not without their pitfalls, they are a potentially valuable tool in teacher education programs aimed at developing reflective thinking. With careful attention to the introduction of the portfolio and guided support throughout the portfolio creation period, many students will invest themselves in the process, enhancing not only their abilities to think reflectively, but also their enthusiasm for learning about themselves, others and the process of teaching (p.78).

Darling-Hammond, Linda, and Jon Snyder. "Authentic Assessment of Teaching in Context." *Teaching and Teacher Education* 16 (2000): 523-545.

Darling-Hammond and Snyder (2000) described and analysed assessment practices in seven teacher training programs in the United States (at Stanford University, Bank Street College, Columbia University, the University of Alaska, Alverno College, University of Southern Maine, the University of California at Santa Barbara) which were considered highly successful on the basis of reputation among scholars and academics; surveys of graduates and employers; and observation of graduates.

They explore the issue of the actual value of the portfolio further. They suggest that portfolios "help make teaching stand still long enough to be examined, shared and learned from" (p.537). They suggest that teaching, and presumably other professional activities, tend to vanish into "amnesia" as the practitioner so busy with daily happenings is unable to recall and reflect on teaching events, achievements and problems. They suggest that it is "through this process of selecting and discussing artefacts of their practice that candidates internalise the standards, examine more deeply what they are doing and what it means and gain multiple perspectives on the meaning of events, thus enhancing their ability to learn from those events Whitford et al. 1999. This notion is implicit in Lee Shulman's (1994) definition of a teaching portfolio as 'the structured documentary history of a (carefully selected) set of coached or mentored accomplishments substantiated by samples of student work and fully realized only through reflective writing, deliberation, and serious conversation' (p.539). Darling-Hammond and Snyder looked at their seven successful programmes and concluded that the benefits of portfolios seemed to relate to their ability to:

- Raise teaching decisions to consciousness and thus make them available for deeper consideration from many perspectives …
- Take a long view of learning and of the development of performance. Because proficient performance must be developed over a long period of time with continuous practice and reflection on practice, a cumulative record help both to scaffold and evaluate that process.
- Support the developmental process by providing benchmarks for good work, vehicles for self-assessment and peer assessment, and opportunities for revision and refinement.
- Connect thinking and performance. This helps to develop the capacity for reflection and action, rather than just one or the other. They bridge the traditional theory-practice divided by asking for evidence of performance along with a discussion of why decisions and actions were taken.
- Provide multiples lenses and multiple sources of evidence on thinking and performance, thus developing many facets of performance and allowing many pathways into learning.
- Make teaching and learning more public, thereby making the development of shared norms and standards possible, as well as making the sharing of knowledge and experience more available.

These factors combine to enhance the candidates' abilities to integrate the knowledge, skills and dispositions required of teaching and to provide tools for continuous development once teaching (p.539).

Darling-Hammond and Snyder found that:

> Where study is both rooted in practice and unrelentingly analytic suggests that the concerns of beginning professionals can move quickly from a focus on self to a focus on students when they have tools to help them train their sights on the effects of their actions and decisions (p.529).

University teacher training portfolios

McLean, Monica, and Joanna Bullard. "Becoming a University Teacher: Evidence from Teaching Portfolios (How Academics Learn to Teach)." *Teacher Development* 4 (2000): 79-101.

McLean and Bullard carried out research on portfolios produced by trainee university teachers in England. The portfolio, together with a departmental assessment, constitute the basis on which a decision is made about whether the probationer should achieve the Keele University Teaching in Higher Education Postgraduate Certificate. The course designers also hope that producing the portfolio will encourage the development of critical enquiry as an integral part of the teaching of the new lecturers (p.82). The course designers hope that production of the portfolio will encourage aspects of teaching such as development of a holistic, rather than a fragmented approach to thinking about teaching. The portfolios consisted of "A selection of materials with an explanatory and critical commentary" (p.100). e.g. teaching plans/lecture notes with self-evaluation; a critical commentary informed by literature; an assessment of and plan for future development (p.100). The research investigated the portfolios of eight probationers. Four were probationary lecturers, three were produced by graduates teaching assistants and one was produced by a research assistant with some teaching responsibilities. The researchers wanted to see if they could find evidence of :

- types of conceptions of teaching
- reflective practice

Six of the eight portfolio holders achieved the teaching certificate and two received awards at a lower level.

The two researchers "read independently all the portfolios with two research questions in mind: what do the portfolios indicate about the group of novice teachers' conceptions of teaching and, as indicators of reflective practice, what evidence is there of a reflective cycle, a self-critical attitude and of engagement with the theory/practice relationship?" (p.86). The researchers used the conceptualisation of concepts of teaching and reflective practice as available in the literature as heuristic devices to evaluate the "thinking and

practices of novice teachers as revealed in the teaching portfolios" (p.84). This was a qualitative investigation. The researchers produced textual evidence from the portfolios, together with interpretation, to indicate probationer characteristics and development. The researchers were aware of the need to probe differences between what practitioners *said* they did and what they might *actually* be doing as evidenced in lesson place and other documents.

The researchers found that "contrary to the characteristics of new teachers outlined in a range of literature, these novice teachers held student-focused conceptions of teaching and made efforts to operationalise them in teaching practices. The new teachers were self-critical, and reviewed and modified their teaching; some reported changed conceptions of teaching" (p. 79). "However, … the extent to which they are able to operationalise these conceptions appears to be mediated by previous experience and by local 'micro-contexts' (Boud and Walker 1998). They stress that such benefits of portfolio assessment can only be realised if the micro, meso and macro context is appropriate and if the purposes and execution of the assessment project is well conceived. The researchers emphasise that an evaluation of portfolios must be contextualised within the intended rationale and aims of the course as a whole and the particular purposes of the portfolio within that context.

The researchers concluded, as a result of finding reflective practice underdeveloped in the probationers, that either the course needed adjustment or that it might take longer than the length of the course for practitioners to become critically reflective and expert (p. 93).

The analysis of portfolios revealed "the difficulties the teachers were having in putting their intentions into practice. This suggests that many of the studies, dealing with conceptions of teaching underestimate the complexity of the conception/practice relationship" (p.92).

This is an intelligent, thoughtful analysis of portfolio content, integrating empirical investigation with well-articulated theoretical conceptualisations of teaching and reflective practice.

As well as throwing light on the use of portfolios in assessment, it suggests means by which portfolios can be used as research resources.

Writing portfolios

Johnston, Brenda. "Some Proposals for Teaching Analytical Writing: A Principled, Holistic, Pedagogic Approach." Ph.D. Thesis. University of Southampton, Southampton, 1998.

In the action research project described in this thesis, I attempted to develop a principled, holistic, pedagogic framework to guide the teaching of analytical writing. I taught academic writing to a class of first year undergraduate students at the American University in Cairo, an English-medium university. The course was assessed by means of portfolio.

I traced the responses of five students to teaching, individual tutorials and peer-assessment over the course of a semester through the medium of five in-depth case studies. I based the case studies on a detailed, qualitative, longitudinal analysis of: their pre-writing notes; various essay drafts; transcripts of individual tutorials and interviews with the students; and products of various research exercises. Because of the level of detail in the data, I was able to trace the nature and extent of student responses to various types of formative assessment, including my inputs, class discussions, and discussions between me and the individual student. The final draft of the major writing assignments was used for summative assessment. Only one of the case study students seemed to find this a barrier to asking questions and entering into exploratory discussions with me, his teacher, at earlier stages of the work. The other case study students, and indeed the other students in the class, took the opportunity to ask questions, to experiment with new writing styles and so on in the formative assessment stages.

I found evidence of uneven development among the students. Sometimes this related to the nature of the teaching input, sometime according to the nature of the assignment and sometimes to individual differences among the students. At times, students sped forward in their learning. At times they seemed to remain stuck at various points and at times when they attempted to use new writing strategies, their writing appeared to regress as they floundered with new concepts and ways of working.

The context in which particular students learned affected their capacities to respond to formative assessment and to make progress. Relevant contextual factors were, for example, the student's educational background and the out-of-class frame of mind of the student.

Broadly speaking, formative assessment worked effectively. A sustained type of formative assessment which worked through various assignments, which each had more than one draft, was made possible by the nature of the portfolio assessment. However, the portfolio itself was a mere passive receptacle of work produced during the course.

One area which was largely unsuccessful was that of peer contributions to formative assessment. The students did not come from an educational background where they were used to contributing actively to their peers' learning in class. They found it hard to reflect on the work of others or to present their views to them.

Medical portfolios

Finlay, I. G., T. S. Maughan, and D. J. T. Webster. "A Randomised Controlled Study of Portfolio Learning in Undergraduate Cancer Education." *Medical Education* 32 (1998): 172-176.

This article is important in that it illustrates the extreme care that has to be exercised when interpreting results of studies on portfolio assessment.

Finlay, Maughan and Webster reported on the use of portfolios in an undergraduate oncology course at the University of Wales College of Medicine, Cardiff. The entire year cohort of 159 students was divided into two groups (study and control groups). The control group did not go through the course described below. "Each student [in the study group] followed a patient with cancer for 9 months, supported by bi-monthly small-group tutorials" (p.172). There were three students per tutorial group. "Each student was allocated one patient with cancer to follow for a minimum of 9 months, during which time he or she visited the patient at home and in hospital and accompanied the patient to outpatients and investigations" (p.173). "Tutorials focused on the students' experiences with their patients, to encourage students in their contact with the patients. Possible treatment options and the predicted course of the cancer were discussed. Students were encouraged to evaluate their own and others' responses to the disease process (p.173). "Tutors were hospital consultants and general practitioners" (p.173). Tutors had support packs "of recent literature in oncology and palliative medicine to facilitate reading" (p.173). "Students kept a portfolio to document their interactions with the patient, a commentary on aspects of the disease, and relevant articles, press cuttings and photographs (Finlay at al, 1994). The format of the portfolio was determined by the student; no formal structure was given" (p.173) "Students recorded triggers to learning and key items" (p.172).

At the end of the course, students' knowledge of management of cancer was assessed on the objective structured clinical examinations (OSCE).[13] The questions were part of a general final examination in Pharmacology and Therapeutics. The students also had to take part in a role play where they explained to a patient's relative the treatment the patient needed. Portfolios were assessed using a standard mark sheet which gave the students marks out of ten on items such as "the context of a patient's disease", "patient's fears and hopes" and "references from the tutor" (p.173), but were not included in the summative assessment. Only 21 portfolios out of a potential group of 80 were submitted at the end of the course.

The researchers found that students in the study showed higher marks in factual knowledge of oncology, especially among the weaker students. Those who had submitted portfolios for assessment had higher score in the OSCE overall than those in the study group who did not submit portfolios. Although the article offers few details, Finlay, Maughan and Webster report that:

> The portfolios were of a remarkably high standard, showing evidence of active learning, triggered by the patient's individual clinical history, the emotional and social effects of cancer on the patient and family, and awareness of the experience of the patient as a consumer of health care services. The portfolios also revealed understanding of factors in clinical decision-making, ethics, palliation and attitudes to death (p.174).

Important points to note here when considering the results of the assessment are that:
(1) probably students who studied on this course did get improved marks, but that was more likely to have been because of the extra course content than the portfolios. After all, only a quarter of the study group actually submitted portfolios. So formative assessment in the shape of tutorials (teacher led formative assessment) and following the case of one patient (reflection led formative assessment) probably helped learning, but the portfolios have been entirely incidental to this process. Also, nearly 60 of the students may not actually have kept portfolios.
(2) most probably those students who returned the portfolios were the most motivated and most capable students, so one would expect their portfolios to show evidence of active learning etc. In the following year, when the entire cohort took this extra oncology course (and 100 out of 160 students submitted portfolios), the overall marks were lower as one would expect, although "they still showed much insight into the impact of cancer on a patient as a person" (p.175). It is unclear how much of this insight is due to the rest of the course content and how much of it developed through the process of producing a portfolio. The researchers do not investigate this.

Snadden, David, and Mary Thomas. "The Use of Portfolio Learning in Medical Education." *Medical Teacher* 20 (1998): 192-199.

This article reviews the use of portfolios in medical education as well as providing a guide to setting up medical portfolios, based on action research,[14] and discussing the need to change current thinking on formal

[13] To remind the reader, Objective Structured Clinical Examination. An OSCE consists of various stations, at each of which, and usually within a few minutes, the candidate must perform a standardised clinical task.

[14] This research is not reported in this article, but can be read about elsewhere. See
Snadden, D., and Thomas, M.L. (1998) Portfolio Learning in General Practice Vocational Training- Does It Work? *Medical Education.*

assessment for portfolios. The main research on which the action research was based took place in Scotland in continuing professional medical education.

The article discusses what portfolios might contain: critical incidents of events with patients; a reflective journal or diary; tutorials and learning plans, and reflection on them; routine clinical experiences; examination preparation material; video recordings of consultations and other relevant material; audits and project work; critical reviews of articles; feedback material; and management material (Snadden and Thomas p.194)

Snadden and Thomas concluded that:

- Not everyone is happy with producing a portfolio in that people have very individual learning styles. For example, young doctors who already have very active learning styles ~~already~~ may perceive portfolio development to be "time wasting or not fitting their needs" (p.198).
- The attitude of the teacher is "fundamental in encouraging and valuing the use and development of a portfolio" (p.198).
- It is essential to be clear about the purpose of the portfolio. If the portfolio is to be used for summative assessment, then students may be reluctant to include information which, for example, illustrate incidents which have not gone well. Such incidents may be "a rich source of learning", but people will be reluctant to include such material if the portfolio is to be included for assessment in a competitive examination (p.196).
- Assessment of portfolios is labour intensive, requiring "careful reading and response to a learner's objectives and evidence of whether they have been met" (p.196). "They are effective as mechanisms to support and facilitate personal learning and growth, but cumbersome in comparative assessments" (p.197). The authors suggest that portfolios "may be best used as additions to assessments to illustrate particular aspects of personal growth that cannot be demonstrated by traditional assessment procedures" (p. 197).

The authors suggest that we have to reconceptualise assessment instead of insisting on "comparing students with each other and with issuing grades or marks … This type of thinking "does not fit easily with portfolios which are essentially non-standardised" (p.197). They conclude that "while the emphasis on grading, excellence and comparison between students and doctors remains in assessment and medicine, it is likely that portfolios will have a greater place as a learning tool than as a summative assessment took, but as a learning tool they may have considerable influence on a learner's performance in summative assessments. If, however, they can be grasped to develop innovative assessments they may have the potential to influence the type of learning in medical curricula to encourage the development of reflective practitioners" (p.197).

The authors review some of the research on portfolio assessment in medical contexts.

- Portfolios are reasonably widely used in nursing education and continuing professional development as well as in general practice training (p.193-240).
- However, there has been little critical appraisal of the effects of this. The main questions on portfolios concern issues of reliability of summative assessment if contents of the portfolios are not standardised; issues of conflict between the use of portfolios for formative and for summative assessment; and, given that these are medical portfolios, issues of privacy and confidentiality regarding the contents of the portfolios.
- In a small interview study of post-registration midwives and their teachers who had used portfolios (Mitchell 1994), the tutors felt portfolios were useful for learning for both themselves and their students. Although some students agreed with this, many felt that the portfolio was not helpful in their learning and not a fair way to assess them. Mitchell suggests that a small sample and the inexperience of both tutors and students may have been a contributory factor in these findings.
- Al-Sheri (1995) reports that portfolios in higher professional training for general practitioners are effective as reflective learning took place, and could be used for postgraduate training purposes.

Davis, M. H., et al. "Portfolio Assessment in Medical Students' Final Examinations." *Medical Teacher* 23 (2001): 357-365.

This article is broad-ranging and not limited to formative assessment. It is an evaluative report on the introduction of portfolios as part of the assessment of final year medical students at Dundee Medical School. It was those who introduced the portfolios who evaluated their implementation. The report investigates issues of reliability, validity, fairness and practicality. The formative aspects of the assessment, which must have taken place, given that this work was discussed and graded throughout the process of portfolio building, are not discussed extensively in this article. Otherwise, the assessment process is carefully described. It would probably be very useful to anyone working in a similar situation (high stakes, competitive, summative assessment), intending to set up a portfolio assessment process.

Snadden, D. et al. Portfolio-based Learning in General Practice Vocational Training. *Medical Education,* 30, pp. 148-152.

Each final year medical student had to produce a portfolio as the second part of their final examination. The portfolio took two years to construct. The first part of the examination consisted of a multiple choice paper; a constructed response paper; and a clinical examination text. The portfolio consisted of the following items. Those marked √ appear to have involved some formative assessment.

- 56 patient presentations which consisted of short summaries of 56 patients seen by the students throughout their course, with reflections on what the students had learned from each encounter and comments on how learning in the 8 presentations completed in Year 5 related to curriculum outcomes.
- A practical procedures card – this was a record, certified by a member of staff, of the 62 practical procedures students had completed or observed throughout the course
- Two pre-registration house officer attachment plans and assessment forms. These forms consisted of grades allocated by supervisors and tutors.
- ✓ 19 case discussions of about 1,500 words each. Each one analysed a patient's history in terms of one of the curriculum's 21 themes. These were marked and feedback given through the course.
- ✓ Reports prepared by the students for their various course units which were graded.
- ✓ Reports written by the students for their elective courses and marked and commented on by members of staff.
- ✓ Fourth-year assignment which was a project marked and commented on by staff.

The assessment process went as follows:

- Students were briefed about the aims and contents of portfolios at the start of Years 4 and 5. An open meeting about the assessment process was held two months before the portfolio examination for staff and students. Written information was distributed to staff, students, internal and external examiners about one month before examination. There were additional verbal briefings for the 25 internal and 8 external examiners.
- *Work carried out before the portfolio review* – e.g. administrative procedures for collecting and recording the content of the portfolios were put in place to avoid disputes over contents etc. Four examiners were allocated to review each portfolio. The four examiners met and discussed the portfolios for which they were responsible and allocated interim grades.
- *Portfolio reviews with the examiners* – Each student was examined by two pairs of examiners on different parts of the portfolio.
- *Work completed after the portfolio review* – this included administrative logging of a total of 22 grades per student.
- The introduction of portfolio assessment was then evaluated through analysis of student results; observer documentation; examiners' evaluation questionnaire; student evaluation questionnaire; and verbal reports from student representatives.

The assessment outcomes were that of the 129 candidates, three withdrew because of illness and 108 students passed, 13 received a conditional pass and five failed (p.359). The results of the portfolio assessment showed low/moderate correlation with the non-portfolio components of the final examination (359-360). The researchers evaluated the portfolio process through questionnaires returned by all 33 examiners and three observers (two internal and one external) commented in writing on portfolio assessment process. The questionnaires used Likert type items plus open ended questions. There was also periodic feedback from the students, and student group discussion with the external examiner after the assessment process was over, written reports from the observers.

The authors of the article reported various conclusions. Not all of these are directly related to formative assessment as this is a wide ranging article.

- The observers came to various conclusions such as that careful administrative procedures had worked well and were necessary to avoid disputes over submitted material, illustrating the high stakes nature of this assessment. The reading burden for internal examiners was high, although they speeded up with practice. The time allocated for the oral examination (50 minutes per student) was valuable for borderline cases, but probably excessive for clear pass cases (p.360).
- The examiners' evaluation questionnaire showed "strong support for the portfolio assessment and its ability to identify the strengths and weaknesses of the students". Some examiners did mention in the open-ended questions that they thought the case study element in the portfolio could be reduced. Some examiners suggested that the results of the other components of the final examination should be included in information they were given. Some thought the patient presentations should be marked before inclusion in the portfolio. Overall the examiners saw the portfolio assessment process as "a robust and successful method of assessing learning outcomes. … It can assess outcomes not easily assessed by other methods; for example, attitudes, personal attributes such as diligence in building the portfolio and aptitude for self-development" (p.363). As such it has considerable intrinsic validity.

- The student evaluation questionnaire, completed after the students knew their results and with an 83% response rate, found that many students expressed reservations about the portfolio assessment. The students were concerned with issues such as excessive paperwork. Other students mentioned a sense of achievement in completing the questionnaire. The students worried about variable standards in marking the case studies and submission of the material to strict deadlines which was stressful. Some were nervous about the novelty of this type of assessment, saying that they felt "less confident as PRHOs [pre-registration house officers] by missing the 'rites of passage' represented by traditional finals" (362).
- On the other hand, the students appreciated that senior doctors took so much care over reading their work and discussing it with them. They thought the assessment was fair and well-organised.
- Student representatives were broadly supportive of portfolio assessment and welcomed the opportunity to present two years of work to senior doctors.
- In terms of reliability, the authors point out that examiners will need further training in applying similar standards in their judgements and that the development of descriptors of grades for each outcome may be helpful.
- In terms of practical implementation, the authors felt that the overall resource required in terms of examiners, time and space may not be "greater than that required for traditional finals in a range of disciplines" (p.364).
- The authors felt that the portfolio assessment process had considerable potential for enhancing student learning, but that this had not been adequately exploited in the first round of portfolio assessment. They planned to address this through greater guidance to both students and staff.

Positive findings about the use of portfolios and formative assessment

Some studies did find that formative assessment of portfolios can enable productive forms of learning to take place. Several of the articles (e.g. Darling 2001; Wade and Yarborough 1996) described a learning environment (individualised learning environment, space to reflect, opportunities for students to develop).

In the case of Darling (2000); Darling-Hammond and Snyder (2000); Johnston (1998); McLean and Bullard 2000; Wade and Yarborough (1996); this happened in the shape of reflection or self-assessment.

In the case of Johnston (1998) this happened in the shape of teacher-led dialogue with the student.

We did not find any instances where peer feedback had helped with formative assessment, but this is most likely to have happened as only one of the articles reviewed used peer assessment and that was in rather unfavourable circumstances. Darling-Hammond and Snyder (2000) did touch on the issue, but not in detail.

Some studies (Darling 2001; Darling-Hammond and Snyder 2000; Davis et al. 2001; and Johnston 1998) did find evidence of the learning process for summative assessment in the portfolios. This assessment was complex and highly subjective, but the authors clearly felt it was possible.

Some studies found evidence of professional growth resulting from the portfolio creation process (e.g. Darling 2001; Darling-Hammond and Snyder 2000; McLean and Bullard 2000)

In some cases, (Johnston 1998) it was formative assessment of the work done that happened to go into a portfolio that appeared to encourage learning.

In other cases (Darling-Hammond and Snyder 2000), it was reflection on the part of the student him/herself and perhaps enhanced by dialogue with the teacher, about the *portfolio contents* and *construction* that enabled the learning to take place. In this case, the portfolio itself was *an active vehicle of learning.*

In the case of Wade and Yarborough 1996, it was unclear whether it was course content and/or the process of portfolio construction that encouraged learning.

Some studies mentioned that producing the portfolios increased the motivation of at least some of the students (Darling 2001; Finlay, Maughan and Webster 1998; Wade and Yarborough 1996).

Some studies discussed the effects of context on student capacities to respond to the portfolio construction process and noted its importance (Johnston 1998; McLean and Bullard 2000).

Some studies discussed the effects of learning style on the ability of the student to respond to the portfolio construction process and found these to be considerable (Johnston 1998; Snadden and Thomas 1998).

None of the articles reviewed investigated the claim made about portfolios that they permit "authentic assessment" which should have greater predictive value about how students will actually perform in professionally related areas, for example.

Problems with formative assessment and portfolios

The studies found a variety of problems, frequently alongside positive findings about portfolios. *Figure 8* offers a description of the problems encountered in implementing portfolio formative assessment.

Figure 8 Problems encountered in the portfolio formative assessment process

Problem	Article
Difficulties in achieving reflection among students	McLean and Bullard (2000); Wade and Yarborough (1996)
Peers unwilling or unable to contribute	Johnston (1998)
Learning style of students not suited to portfolios	Snadden and Thomas (1998)
Feeling of missing out on a rite of passage by producing a portfolio instead of doing traditional examinations	Davis et al. (2001)
Confusion and anxiety about producing a portfolio	Darling (2001); Wade and Yarborough (1996)
Over-concern about format of the portfolio	Darling (2001)
Inappropriate or inadequate feedback	Wade and Yarborough (1996)
Conflict between summative and formative assessment	Snadden and Thomas (1998)

Some of these problems seemed to relate to inadequate preparation on the part of the portfolio assessors and/or an over-free hand for the students. Some students were unclear about what expected of them. They sometimes lacked models and previous experience of portfolios and became very worried about the nature of the task (Darling 2001; Wade and Yarborough 1996). As Wade and Yarborough pointed out, such initial worry is natural and probably indicative of personal growth, but continuing anxiety is likely to be a sign that the process is not working. This relates to the point made earlier in the section *Implications of theory for teaching, learning and assessment* p.67-68 that learning environments should be carefully structured and scaffolded in order to be productive.

Frequently, some students in a group could cope with the processes involved in producing a portfolio while others could not. This indicates individual differences in learning styles, but also probably different earlier educational and other experiences which may or may not have prepared the student for the portfolio task. Cultural differences in educational experiences should also be recognised. For example, Johnston (1998) reported that students in her study, used to rote learning and an authoritarian classroom environment, found it hard to participate meaningfully in peer formative assessment, even given careful guidance. Coming from another angle, Snadden and Thomas (1998) found that young doctors who already learned in a very active way perceived portfolios as a waste of time. Davis et al. (2001) found that medical students felt somewhat deprived of their traditional finals "rite of passage" by the portfolio system. Student attitudes to portfolios are always contextualised.

Some students in some of the studies seemed to become rather pre-occupied with producing an exciting and novel portfolio, rather than the everyday hard work and reflection that presumably most portfolios in this setting should consist of. One can also imagine as Darling-Hammond and Snyder (2000) mention that in these situations the portfolios can "privilege a candidate's ability to select and write about artefacts to teaching more than the candidate's capacity actually to teach well in the classroom in the heat of real situations with real students" (p.539). Darling-Hammond and Snyder then argue that is important that assessors be expert and aware enough to assess whether they are being presented with solid evidence about teaching experiences and capacities or a public relations exercise.

At times, it seemed that students found it hard to become reflective. McLean and Bullard (2000) found this. The researchers concluded, that either the course needed adjustment or that it was not long enough for the students to develop their reflective capacities. They also found that the students had difficulties in putting their ideas into practice. One should probably not blame the nature of portfolios and formative assessment for this. Rather these processes are complex and control over them probably takes a long time to achieve. In some cases, environments may just not be conducive to reflection developing. One can imagine inhibiting contexts such as overly authoritarian teaching or an over-full syllabus. None of the studies reviewed seemed to come into this category.

There may be a conflict if portfolios are used for both formative assessment and summative assessment. If portfolios are to be used for summative assessment, students may be reluctant to include work which would show them as being anything less than competent. (See e.g. Snadden and Thomas 1998, p.196). However, other studies have found that both types of assessment can be included as part of a portfolio. Johnston (1998) found, for example, that both could be included as part of two stage assignments.

How to make formative assessment in portfolios work better

In order to lessen anxiety about portfolios, Darling (2001) suggested:

- asking students to help generate ideas for specific content helps to bring them into the process early and proactively ..
- negotiating evaluative criteria helps students become clearer about the broader purposes for constructing portfolios and the goods associated with them, including the internal good of taking ownership of one's own learning. Working through this process is also a good way to link portfolio evaluation with the sorts of assessment practices they will engage in as elementary teachers (Meyer and Tusin 1999). Teachers at all levels struggle to construct fair and meaningful standards for evaluating students' work. Portfolios offer special, complex challenges because they attempt to capture so many aspects of learning. Constructing standards together was one way of acknowledging the importance of student autonomy with regard to the project. "What knowledge was most important for your own growth?" is one question we work at keeping in the foreground (p.118).

It is perhaps important to note that to implement this advice could prove very intimidating for those from a restricted educational background and/or perhaps those who are young and (as yet) unused to taking responsibility for their own learning. Context is all important.

In order to help with structuring portfolios by offering a balance between guidance and personal choice, Darling (2001) suggested providing some scaffolding with sample portfolios. "We decided students should feel free enough to pursue independent directions, but secure enough to fall back on an established framework if it was needed" (Darling p.118-9). This seems like a very sensible idea for ensuring the students have some idea of what is actually expected of them. Again not all portfolio assessors may wish their students to have such a degree of freedom.

In order to provide in-course support for portfolios to students unsure about how to construct them, Darling suggested providing class discussion sessions, small study groups with weekly meetings built into the timetable, handing in draft parts of the portfolios to instructors as well as peers for feedback (p.119). This sounds like another very sensible idea, but it may be subject to resource constraints, although then the issue of what to prioritise raises its head.

Wade and Yarborough (1996) suggested that in order for portfolios to work as reflective instruments, students had to be prepared to invest effort in the portfolios which was more likely to happen when they

- understood the value and nature of reflection
- wanted to learn to do their best
- wanted to receive good grades
- enjoyed the overall course and found it a developmental experience
- received a tool in the shape of a portfolio which they could use when seeking employment
- were able to make connections between their outside experiences and the portfolio
- the nature of required assignments assisted reflection (p.77).

The researchers therefore suggest that these issues are attended to in course design with care taken, as far as possible, to encourage these different types of motivation and structure (p.77-78).

As far as is possible and practical, individual differences should be recognised. Support will be especially necessary for students who find relatively unstructured learning and assessment environments difficult to cope with.

It seems likely that as portfolios become embedded within any particular course, their path will run more smoothly. Staff will have greater experience of using them. Students will have expectations of having to produce one and will have a tradition to call on in the shape of models and word of mouth experiences. Administrations systems will have been worked out.

Conclusions on the links between learning, formative assessment and portfolios

In general, when carefully and sensitively implemented with due regard for the local context, the opportunities offered by portfolios for learning and various kinds of formative assessment are considerable. They may be particularly appropriate for particular areas such as professional training and other areas where there is a strong relationship between theory and practice. However, it is hard to imagine an area which would not benefit from students collecting work over a period of time and reflecting on it. Portfolios need only be part of the assessment arrangements for any one course.

It should perhaps be noted that portfolios are not an answer to all ills. Some students may find them hard to work with for one or more reasons as listed above in the section *Problems with formative assessment and portfolios* p.80-81. Care has to be taken to minimise such problems.

There does seem to be at least one major gap in the research in this area: predictive validity. Although portfolios seem to offer much promise for improving practice and that anecdotal evidence supports this, there is little systematic research evidence connecting portfolio achievements with future achievements. This would be complex research to carry out but Darling-Hammond and Snyder (2000) suggest, in the case of teacher training at least, that it would be possible. Darling-Hammond and Snyder suggest a series of research questions, related to predictive validity and teacher development which it would be useful to ask in the field of teacher training. The questions are:

- How well do different types of assessments measure the capacity to teach? What evidence can be developed of the predictive and consequential validity of various measures?
- What are the effects on teacher learning of the use of different types of assessment?
- Given that no single measure of teaching is adequate to the task of representing such a complex activity, what mix of assessment methods, instruments, and sources of evidence seem to provide the greatest leverage on teacher development, on the one hand, and valid assessment on the other? (p.543)

These questions could easily be transferred to other fields.

Section Four The practicality of using portfolios for assessment

I had ambitions at one stage in this project of writing a review of the practicality of portfolio assessment. In the light of deadlines looming, this ambition has bitten the dust. However, I had been collecting references on practicality issues and any readers contemplating introducing portfolio assessment might wish to follow up these references listed in *Figure 9* below..

This section focuses on technically practical matters, rather than the pedagogically related issues of previous sections. It is important to consider practical issues at the macro, meso and micro levels since co-ordination and planning at all these levels is necessary for assessment to work effectively as Yorke (1998) points out. He makes various general recommendations at the institutional level about the management of assessment as follows:

- The institutional system for assessment should be based on a thorough functional analysis of who is expected to do what, and why. Critical issues are whether the functions articulate coherently; whether there are any 'gaps' or duplications in the system; and whether there are any assignments of task to an inappropriate level of member of staff.
- The system should be audited for operational strengths and weaknesses, with the intention of rectifying any detected weaknesses, and of spreading desirable practice throughout. The questions used by HEQC's quality auditors are helpful in this regard (HEQC, 1995).
  - What are you trying to do?
  - Why are you trying to do it?
  - How are you doing it?
  - Why are you doing it that way?
  - Why do you think that is the best way of doing it?
  - How do you know it works?
  - How do you improve it?
- Assessment needs to be given a higher profile, both conceptually and technically within curriculum design and implementation.
- A continuing series of staff development activities is needed, for academic and support staff, as appropriate, in respect of needs identified as a result of functional analysis.

- There is an ongoing need to give consideration to the conceptual, structural and temporal relationships of assessment to the curriculum, with a view to ensuring that assessment can be used to maximum advantage in respect of the purposes ascribed to it (p.114).

These recommendations could easily be applied to the implementation of portfolio schemes. The LTSN generic centre has also published a series of guides on assessment for people at different levels in higher education: senior managers (Yorke 2001; heads of department Mutch and Brown 2001; lecturers Brown 2001; and students Race 2001). Comments from these will be helpful in the implementation of portfolio assessment or indeed evaluation of other forms of assessment.

Many studies include at least some comments on portfolio assessment. They often differ widely in their comments and assessment. I shall summarise the information I have collected so far in *Figure 9* below. This list is not exhaustive. The issues covered in *Figure 9* often (although not always) include problems encountered, but these are often mentioned in the context of how they were overcome.

It is perhaps important to note my general impression that although practical issues are mentioned frequently, they are rarely investigated empirically at a strategic, co-ordinated level in detail. They tend to be mentioned as part of research articles which are focused primarily on other issues.

Figure 9 Practicality of using portfolios

Issue addressed	Study
Need for a co-ordinated management and funding system	Darling-Hammond and Snyder (2000); Underwood 1998
Resource required compared to traditional means of assessment	Davis et al. (2001)
Financial costs of marking portfolios	Reckase (1995); LeMahieu, Gitomer and Eresh (1995); Mehrens 1992
Excessive paperwork for students	Davis et al. (2001)
Sustained construction of portfolios among professionals for professional development	Smith and Tillema (2001)
Making sure portfolios contain all relevant pieces of work in a large scale assessment scheme	Supovitz, MacGowan and Slattery (1997)
Staff training	Payne et al. 1993
Benefits and burdens for teachers in schools	Stecher (1998)
Time needed for reading portfolios	Davis et al. (2001)
Whose work is being assessed in a portfolio – role of peers, teachers and parents	Gearhart and Herman 1998
Piloting of portfolio assessment	Finlay, Maughan and Webster 1998; Payne at al. 1993
Practical embedding of portfolio assessment within curriculum and timetable	Davis et al. 2001; Finlay, Maughan and Webster 1998; Payne et al. 1993
Oral examinations and portfolio assessment	Davis et al. (2001)

To give an example of the kind of the comments studies make, I shall report on what two studies said about the practical implication of introducing portfolio assessment. A wide ranging article such as Davis et al. (2001), notes practical issues as follows. The observers came to various conclusions such as that careful administrative procedures had worked well and were necessary to avoid disputes over submitted material, illustrating the high stakes nature of this assessment. “The reading burden for internal examiners was high”, although they speeded up with practice. “The time allocated for the oral examination (50 minutes per student) was valuable for poor/borderline cases, but probably excessive for ‘safe pass’ and above cases” (Davis et al. p.360). The students were concerned with issues such as excessive paperwork.(p.362). In terms of practical implementation, the authors felt that the overall resource required in terms of examiners, time and space may not be “greater than that required for traditional finals in a range of disciplines” (p.364).

To take another example, Payne et al. (1993) in their account of introducing portfolios into an integrated engineering course at Sheffield Hallam University make similar comments. They note that “ …during the first year [of the course] … the students had difficulty understanding the concept of a portfolio and did not gather appropriate material for it” (p.38). So the course developers decided that a supporting Professional and Personal Development (PPD) programme was necessary. They report in detail about how the course was set up:

- As part of the PPD course, the students to “gather[ed], evaluate[d], review[ed] and present[ed] relevant material for their portfolio” (p.39). Students in groups of 12-20 had weekly 1½ hour sessions led by a

member of staff from the Engineering school. In addition, PPD tutors saw students for 10-20 minutes, three times each year "to discuss progress and make action plans" (p.40).
- Five core staff were responsible for developing the PPD programme and met regularly. 15 staff taught on the PPD course.
- All first and second year engineering students were later included in the portfolio assessment scheme. 20 tutors, many new to experiential, facilitative teaching, were involved in assessing the portfolios.
- At the beginning of the portfolio implementation, portfolios were merely seen "as an alternative way of assessing learning" (p.41). However, the process of learning became very open to view and the course leaders decided that they had to be very clear about what was being assessed. They decided the portfolio would assess "as far as possible in a job or work-related context:
    - the application of engineering knowledge and skills
    - the professional functioning of a potential professional engineer" (p.41).

They noted problems such as because there was not any significant assessment of the PPD course and portfolios in the first years of the degree, students had tended to miss a large proportion of the classes. They made suggestions such as the following:

- It is important that core staff are give time to "argue through, draft and evaluate the practical aspects of the course" (p.40).
- Staff involved in implementation of portfolio production among students need "intensive training …. Otherwise the result will be at best patchy, at worst disastrous" (p.40).
- Bringing about changes in the assessment process through portfolios helped bring about other developments in the engineering course as the process of portfolio assessment is a visible process and also because portfolio assessment in engineering was new so there were no experts to consult or traditions to follow.

It is important to pilot development on a small group and then to move to persuading other staff and students that portfolio assessment is "better, fairer and workable" (p.42).

Overall conclusions

For specific conclusions, gaps in research and recommendations on the issues discussed, it is best to consult the relevant sections earlier in the report. In this concluding section, we would like to draw together some of the issues discussed at a more general level. The comments here can be applied to various types of assessment, apart from portfolios.

All assessment has various integral aspects. Each assessment implemented:

- embodies educational and philosophical values (attitudes to knowledge and ways of knowing; attitudes to learning). These may be conscious or unconscious. They are made manifest in highly practical ways such as the means by which assessment judgements are reached and formative assessment viewed.
- is part of a wider context of approaches to teaching, learning and assessment. This context operates at the national, institutional, disciplinary, departmental, individual classroom and individual teacher and student level.
- each assessment is also takes place within particular resource possibilities and constraints. These resource possibilities will be shaped by actual financial resources available, but also by prioritisation of funds

The context to assessment may be strategically co-ordinated or it may not. Assessment aims and practices may be aligned with teaching and learning approaches or they may not. Assessment aims and practices may be aligned with the fundamental educational and philosophical values of individual assessors and students or they may not. These alignments may be intentional or they may be a result of historical accident.

One final word concerns gaps in existing research. We have discussed gaps in research as it relates to the issues discussed in this report. However, there are some gaps in other areas. For example, research on assessment would probably benefit from more "qualitative and ethnographic validation procedures- like interviews, observations, and thick descriptions to understand the role an assessment plays within a specific programme or institution" (Huot 1996, p.561-2).

Another example, is whether particular forms of assessment such as portfolio assessment advantage or disadvantage particular groups and if so how does this work (Herman and Winters 1994, p.50). It has been suggested that portfolios in schools might advantage middle class children with greater home resources to call on. It has also been suggested in one study that girls at school and certain ethnic groups tended to

score higher then boys and other ethnic groups (LeMahieu, Gitomer and Eresh 1995, p.15-16). Are there parallel differences in higher education and if so are these causes for concern?

List of references

Adler, S. "Reflective Practice and Teacher Education." *Reflective Practice in Social Studies*. Ed. E. W. Ross. Vol. Bulletin 88. Washington: National Council for the Social Studies, 1994. 51-58.

Allen, Michael. "Valuing Differences: Portnet's First Year." *Assessing Writing* 2 (1995): 67-89.

Al-Sheri, A. "Learning by Reflection in General Practice: A Study Report." *Education for General Practice* 7 (1995): 237-248.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: Authors, 1985.

Applebee, A., J. A. Langer, and I. V. S. Mullis. *The Writing Report Card, 1984-1988: Findings from the Nation's Report Card*. Princeton, NJ: National Assessment of Educational Progress, 1990.

Athanases, S. Z. "Teachers' Reports of the Effects of Preparing Portfolios of Literacy Instruction." *Elementary School Journal* 94 (1994): 421-439.

Bakhtin, Mikhail, and P. N. Medvedev. *The Formal Method in Literary Scholarship: A Critical Introduction to Sociological Poetics*. Translated by Albert J Wehrle ed. Cambridge: Harvard University Press, 1985.

Bakhtin, Mikhail. *Speech Genres and Other Late Essays*. Ed. Caryl Emerson and Michael Holquist. Trans. V.W. McGee. Austin: University of Texas Press, 1986.

Barton, J., and A. Collins. "Portfolios in Teacher Education." *Journal of Teacher Education* 44 (1993): 200-210.

Baume, David, and Mantz Yorke. "The Reliability of Assessment by Portfolio on a Course to Develop and Accredit Teachers in Higher Education." *Studies in Higher Education* (In press for 2002).

Baume, David. A Briefing on Assessment of Portfolios. LTSN Generic Centre. Assessment Series, 2001.

Becker, Howard, Blanche Geer, and Everett Hughes. *Making the Grade: The Academic Side of College Life*. New York: John Wiley and Sons, 1968.

Benhabib, S. *Situating the Self: Gender, Community and Postmodernism in Contemporary Ethics*. New York: Routledge, 1992.

Biddle, J. R. Portfolio Development in Teacher Education and Educational Leadership. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, San Antonio, TX.

Biddle, J. R., and T. J. Lasley. Portfolios and the Process of Teacher Education. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, 1992.

Black, P., and D. Wiliam. "Assessment and Classroom Learning." *Assessment in Education* 5 (1998): 7-74.

Borg, Walter R., and Meredith Demien Gall. *Educational Research: An Introduction*. New York: Longman, 1989.

Borko, H., et al. "Student Teaching Portfolios: A Tool for Promoting Reflective Practice." *Journal of Teaching Education* 48 (1997): 347-357.

Brandt, Deborah. "The cognitive as the social: an ethnomethodological approach to writing process research." *Written Communication* 9.3 (July 1992): 315-355.

Britton, J., et al. *The Development of Writing Abilities*. London: Macmillan, 1975.

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

Broadfoot, Patricia. "Records of Achievement and the Learning Society: A Tale of Two Discourses." *Assessment in Education* 5 (1998): 447-477.

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

Brown, George, and Gillian Yule. *Discourse Analysis*. New York: Cambridge UP, 1983.

Brown, George. Assessment: A Guide for Lecturers. LTSN Generic Centre. Assessment Series. 2001.

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

Bruner, J. *The relevance of education*. London: Allen and Unwin, 1972.

Bruner, J. S. "The Process of Education Revisited." *Phi Delta Kappan* 52 (1971): 18-21.

Bruner, Jerome. "Vygotsky: A Historical and Conceptual Perspective." *Culture, Communication and Cognition: Vygotskian Perspectives*. Ed. J. Wertsch. Cambridge: Cambridge University Press, 1985.

Callahan, S. "Tests Worth Taking? Using Portfolios for Accountability in Kentucky." *Research in the Teaching of English* 31 (1997): 295-336.

Centra, John A. "The Use of the Teaching Portfolio and Student Evaluations for Summative Evaluation." *Journal of Higher Education* 65 (1994): 555-570.

Challis, Maggie. "AMEE Medical Education Guide No.11 (revised): Portfolio-Based Learning and Assessment in Medical Education." *Medical Teacher* 21 (1999): 370-386.

Cope, Bill, and Mary Kalantzis, eds. *The Powers of Literacy: A Genre Approach to Teaching Writing*. London: Falmer Press, 1993.

Corbett, H. D., and B. L. Wilson. *Testing, Reform and Rebellion*. Norwood, NJ: Ablex, 1991.

Crocker, L., and J. Algina. *Introduction to Classical and Modern Test Theory*. Fort Worth, TX: Rinehart and Winston, 1986.

Cronbach, L. J. "Five Perspectives on Validity Argument." *Test Validity*. Ed. H. Wainer. Hillsdale, NJ: Erlbaum, 1988.

Cronbach, L. J. *Essentials of Psychological Testing*. 5th ed. New York: Harper and Row, 1990.

Cronbach, L. J. "Test Validation." *Educational Measurement*. Ed. R. L. Thorndike. 2nd ed. Washington, DC: American Council on Education, 1971.

Cruickshank, D. R. *Reflective Teaching: The Preparation of Students of Teaching*. Reston: Association of Teacher Educators, 1987.

CVCP, DfEE, and HEQE. Skills Development in Higher Education, 1998.

Darling, L. Farr. "Portfolio as Practice: The Narratives of Emerging Teachers." *Teaching and Teacher Education* 17 (2001): 107-121.

Darling-Hammond, Linda, and Jon Snyder. "Authentic Assessment of Teaching in Context." *Teaching and Teacher Education* 16 (2000): 523-545.

Davis, M. H., et al. "Portfolio Assessment in Medical Students' Final Examinations." *Medical Teacher* 23 (2001): 357-365.

De Beaugrande, Robert, and Wolfgang Dressler. *Introduction to Text Linguistics*. New York: Longman, 1981.

Delandshere, G., and A. R. Petrosky. "Capturing Teachers' Knowledge: Performance Assessment a) And Post-Structuralist Epistemology; b) From a Post-Structuralist Perspective, c) and Post-Structuralism. d) None of the Above." *Educational Researcher* 23 (1994): 11-18.

Department for Education and Employment, and Northern Ireland Department of Education. Targets for our Future. London, DFEE, 1997.

DeRemer, Mary. "Writing Assessment: Raters' Elaboration of the Rating Task." *Assessing Writing* 5 (1998): 7-29.

Diederich, Paul, John W. French, and Sydell T. Carlton. Factors in Judgements of Writing Quality. Princeton Educational Testing ServiceRB No. 61-15, ERIC ED 002 172.

Duranti, Alessandro, and Charles Goodwin, eds. *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press, 1992.

Editorial. "Editorial." *Assessment in Education* 5 (1998): 301-307.

Eisner, E. "Forms of Understanding and the Future of Educational Research." *Educational Researcher* 22 (1993): 5-11.

Elton, Lewis, and Diana Laurillard. "Trends in Research on Student Learning." *Studies in Higher Education* 4 (1979): 87-102.

Elton, Lewis. *Teaching in Higher Education*. London: Kogan Page, 1987.

Falchikov, Nancy, and Judy Goldfinch. "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks." *Review of Educational Research* 70 (2000): 287-322.

Feldt, L. S., and R. L. Brennan. "Reliability." *Educational Measurement*. Ed. R. L. Linn. 3rd ed. Washington: The American Council on Education and the National Council on Measurement in Education, 1989.

Fenstermacher, G. "The Concepts of Method and Manner in Teaching." *Effective and Responsible Teaching*. Ed. A. D. F. Oser and J. L. Patry. San Francisco, CA: Jossey-Bass, 1992. 95-108.

Feyerabend, Paul. *Science in a Free Society*. London: NLB, 1978.

Finlay, I. G., T. S. Maughan, and D. J. T. Webster. "A Randomised Controlled Study of Portfolio Learning in Undergraduate Cancer Education." *Medical Education* 32 (1998): 172-176.

Fish, Stanley. *Is There a Text in this Class? The Authority of Interpretive Communities*. Cambridge, Mass: Harvard University Press, 1980.

Ford, M. P., and M. M. Ohlhausen. Portfolio Assessment in Teacher Education Courses: Impact on Students' Beliefs, Attitudes and Habits. Paper presented at the annual meeting of the National Reading Conference, Palm Springs, CA.

Foucault, Michel. "The Order of Discourse." *Untying the Text: A Post-Structuralist Reader*. Ed. R. Young. Boston: Routledge and Kegan Paul, 1981. 48-78.

Foucault, Michel. *Power/Knowledge: Selected Interviews and Other Writings 1972-1977*. ed. C. Gordon. Trans. C. Gordon, L. Marshall, J. Mepham, & K. Soper. New York: Pantheon, 1980.

Fourali, Chahid. "Using Fuzzy Logic in Educational Measurement: The Case of Portfolio Assessment." *Evaluation and Research in Education* 11 (1997): 129-148.

Frederiksen, J. R., and A. Collins. "A Systems Approach to Educational Testing." *Educational Researcher* 18 (1989): 27-32.

Frederiksen, John R., Mike Spiusic, and Miriam Sherin. "Video Portfolio Assessment: Creating a Framework for Viewing the Functions of Teaching." *Educational Assessment* 5 (1998): 225-297.

Gearhart, Maryl, and Joan Herman. "Portfolio Assessment: Whose Work Is It? Issues in the Use of Classroom Assignments for Accountability." *Educational Assessment* 5 (1998): 41-55.

Gearhart, Maryl, and Shelby A. Wolf. "Issues in Portfolio Assessment: Assessing Writing Processes from Their Products." *Educational Assessment* 4 (1997): 265-296.

Geerzt, Clifford. "Thick Description: Towards and Interpretive Theory of Culture." *The Interpretation of Culture*. Ed. Clifford Geertz. New York: Basic Books, 1973.

Gitomer, D. H. "Performance Assessment and Educational Measurement." *Construction versus Choice in Cognitive Measurement*. Ed. R. E. Bennet and W. C. Ward. Hillsdale: Erlbaum, 1993. 241-63.

Guba, Egon G., and Yvonna Lincoln. *Fourth Generation Evaluation*. London: Sage, 1989.

Halliday, Michael. *Language as Social Semiotic*. Baltimore: Arnold, 1978.

Hartog, P., and E. C. Rhodes. . *The Marks of Examiners*. London: Macmillan, 1936.

Hartog, P., and E. C. Rhodes. *An Examination of Examinations*. London: Macmillan, 1935.

Heller, Joan I., Karen Sheingold, and Carol M. Myford. "Reasoning About Evidence in Portfolios: Cognitive Foundational for Valid and Reliable Assessment." *Educational Assessment* (1998): 5-40.

HEQC. Notes for the Guidance of Auditors. London: HEQC, 1995.

Herman, Joan L., Maryl Gearhart, and Eva L. Baker. "Assessing Writing Portfolios: Issues in the Validity and Meaning of Scores." *Educational Assessment* 1 (1993): 201-224.

Herman, Joan, and Lynn Winters. "Portfolio Research: A Slim Collection." *Educational Leadership* (1994): 48-55.

Heywood, John. *Assessment in Higher Education: Student Learning, Teaching, Programmes and Institutions*. London: Jessica Kingsley, 2000.

Hinett, K. "Developing Student Learning through the Use of Self-assessment in Higher Education." Dissertation. University of Central Lancashire, 1997.

Huot, B. "The Influence of Holistic Scoring Procedures on Reading and Rating Student Writing." *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Ed. M. Williamson and B. Huot. Cresskill, NJ: Hampton Press Inc, 1993. 206-36.

Huot, Brian. "Towards a New Theory of Writing Assessment." *College Composition and Communication* 47 (1996): 549-566.

James, D. Marking the Graduate: Assessment Events as Social Practices. Paper presented at British Educational Research Association Annual Conference, York, September, 1997.

Johnston, Brenda. "Some Proposals for Teaching Analytical Writing: A Principled, Holistic, Pedagogic Approach." Dissertation. University of Southampton, Southampton, 1998.

Johnston, P. H., P. Weiss, and P. Afflerbach. Teachers' Evaluation of the Teaching and Learning in Literacy and Literature. Report Series 3.4. Albany: State University of New York at Albany, Center for the Learning and Teaching of Literature.

Knight, Peter. A Briefing on Key Concepts: Formative and Summative, Criterion and Norm-referenced Assessment. LTSN Generic Centre: Assessment Series, 2001.

Knowles, M. "Andragogy: An Emerging Technology for Adult Learning." *Education for Adults: Adult Learning and Education*. Ed. M. Tight. London: Croom Helm, 1970.

Kolb, D. A. *Experiential Learning*. Chicago: Prentice Hall, 1984.

Kolb, David. "The Process of Experiential Learning." *Culture and Processes of Adult Learning*. Ed. Mary Thorpe, Richard Edwards, and Ann Hanson. London and New York: Routledge, 1993. 138-56.

Koretz, D., B. Stecher, and E. Deibert. The Reliability of Scores from the 1992 Vermont Portfolio Assessment Program. (Tech. Rep. No. 355). Los Angeles: University of California CRESST; Center for the Study of Evaluation.

Koretz, Daniel, et al. "The Vermont Portfolio Assessment Programme: Findings and Implications." *Educational Measurement: Issues and Practice* (1994): 5-16.

Koretz, Daniel. "Large-Scale Portfolio Assessments in the US: Evidence Pertaining to the Quality of Measurement." *Assessment in Education* 5 (1998): 309-334.

Labov, William, ed. *Location Language in Time and Space*. New York: Academic, 1980.

LeMahieu, P., D. H. Gitomer, and J. T. Eresh. Portfolios in Large-Scale Assessment: Difficult but not Impossible. Unpublished Manuscript, University of Delaware, 1993.

LeMahieu, Paul G., Drew H. Gitomer, and JoAnne Eresh. "Portfolios in Large-Scale Assessment: Difficult But Not Impossible." *Educational Measurement: Issues and Practice* (1995): 11-28.

Levinson, Stephen C. *Pragmatics*. New York: Cambridge UP, 1983.

Lindstrom, Lamont. "Context Contests: Debatable Truth Statements on Tanna (Vanuata)." *Rethinking Context: Language as an Interactive Phenomenon*. Ed. Alessandro Duranti and Charles Goodwin. Cambridge: Cambridge University Press, 1992. 101-24.

Linn, R. L., E. L. Baker, and S. B. Dunbar. "Complex Performance-based Assessment: Expectations and Validation Criteria." *Educational Researcher* 20 (1991): 5-21.

Loughran, John, and Deborah Corrigan. "Teaching Portfolios: A Strategy for Developing Learning and Teaching in Preservice Education." *Teaching and Teacher Education* 11 (1995): 565-577.

MacIntyre, A. *After Virtue*. Note Dame, IN: Notre Dame University Press, 1984.

Madaus, C. "The Influence of Testing on the Curriculum." *Critical Issues in Curriculum: 87th Yearbook of the National Society for the Study of Education*. Ed. L. Tanner. Chicago: University of Chicago Press, 1988. 83-121.

Marton, F., and R. Saljo. "Approaches to Learning." *The Experience of Learning*. Ed. F. Marton, D. Hounsell, and N. J. Entwistle. Edinburgh: Scottish Academic Press, 1984.

Mathers, N. J., et al. "Portfolios in Continuing Medical Education - Effective and Efficient." *Medical Education* 33 (1999): 521-530.

McCarthey, Sarah, and Taffy E. Raphael. "Alterative Research Perspectives." *Reading/Writing Connections: Learning from Research*. Ed. Judith W. Irwin and Mary Anne Doyle. Newark, Delaware: International Reading Association, 1992. 2-30.

McCormick, Kathleen. "The Cultural Imperatives Underlying Cognitive Acts." *Reading-To-Write: Exploring a Cognitive and Social Process*. Ed. Linda Flower, et al. New York: Oxford University Press, 1990. 194-218.

McLean, Monica, and Joanna Bullard. "Becoming a University Teacher: Evidence from Teaching Portfolios (How Academics Learn to Teach)." *Teacher Development* 4 (2000): 79-101.

Mehrens, W. A. "Using Performance Assessment for Accountability Purposes." *Educational Measurement: Issues and Practice* 11 (1992): 3-9.

Mercer, Neil. "Language and the Guided Construction of Knowledge." *Language and Education*. Ed. G. Blue and R. Mitchell. Clevedon: Multilingual Matters, 1996. 28-40.

Messick, S. The Interplay of Evidence and Consequences in the Validation of Performance Assessments. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1992.

Messick, Samuel. "Meaning and Values in Test Validation: The Science and Ethics of Assessment." *Educational Researcher* 23 (1989): 5-12.

Meyer, D., and L. Tusin. "Preservice Teachers' Perceptions of Portfolios: Process versus Product." *Journal of Teacher Education* 50 (1999): 131-139.

Mezirow, J. "A Critical Theory of Adult Learning and Education." *Adult Learning and Education*. Ed. M. Tight. London: Croom Helm, 1981.

Miller, M. David, and Sue M. Legg. "Alternative Assessment in a High-Stakes Environment." *Educational Measurement: Issues and Practice* (1993): 9-15.

Milton, G., H. R. Pollio, and J. A. Eison. *Making Sense of College Grades*. San Francisco: Jossey-Bass, 1986.

Mishler, E. G. "Validation in Inquiry-Guided Research." *Harvard Educational Review* 60 (1990): 415-442.

Mitchell, M. "The Views of Students and Teachers on the Use of Portfolios as a Learning and Assessment Tool in Midwifery Education." *Nurse Education Today* 14 (1994): 38-43.

Moss, Pamela. "Can There Be Validity Without Reliability?" *Educational Researcher* 23 (1994): 5-12.

Moss, Patricia. "Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment." *Review of Educational Research* 62 (1992): 229-258.

Murphy, Sandra, Jan Bergamini, and Paul Rooney. "The Impact of Large-Scale Assessment Programmes on Classroom Practice: Case Studies of the New Standards Field-Trial Portfolio." *Educational Assessment* 4 (1997): 297-333.

Murphy, Sandra. "Portfolios and Curriculum Reform: Patterns in Practice." *Assessing Writing* 1 (1994): 175-206.

Mutch, Alistair, and George Brown. Assessment: A Guide for Heads of Department. LTSN Generic Centre. Assessment Series. 2001.

Newmann, F. M. "Higher Order Thinking in Teaching Social Studies: A Rationale for the Assessment of Classroom Thoughtfulness." *Journal of Curriculum Studies* 22 (1990): 41-56.

Nystrand, Martin, Allan S. Cohen, and Norca M. Dowling. "Addressing Reliability Problems in the Portfolio Assessment of College Writing." *Educational Assessment* 1 (1993): 53-70.

Nystrand, Martin, Stuart Greene, and Jeffrey Wiemelt. "Where Did Composition Studies Come From? An Intellectual History." *Written Communication* 10.3 (July 1993): 267-333.

Nystrand, Martin. "A Social-Interactive Model of Writing." *Written Communication* 6.1 (January 1989): 66-85.

Paulson, F. L., P. P. Paulson, and C. A. Meyer. "What Makes a Portfolio a Portfolio?" *Educational Leadership* 48 (1991): 60-63.

Payne, Roger N., et al. "Portfolio Assessment in Practice in Engineering." *International Journal of Technology and Design Education* 3 (1993): 37-42.

Perkins, D. N., Eileen Jay, and Shari Tishman. "Beyond Abilities: A Dispositional Theory of Thinking." *Merrill-Palmer Quarterly* 39.1 (January 1993): 1-21.

Pitts, John, Colin Coles, and Peter Thomas. "Educational Portfolios in the Assessment of General Practice Trainers: Reliability of Assessors." *Medical Education* 33 (1999): 515-520.

Pitts, John, Colin Coles, and Peter Thomas. "Enhancing Reliability in Portfolio Assessment: 'Shaping' the Portfolio." *Medical Teacher* 23 (2001): 351-355.

Polyani, M. *Personal Knowledge: Towards a Postcritical Philosophy*. Chicago: University of Chicago Press, 1962.

Posner, G. I. *Analysing the Curriculum*. New York: McGraw-Hill, 1992.

Pula, J., and B. Huot. "A Model of Background Influences on Holistic Raters." *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Ed. M. Williamson and B. Huot. Cresskill, NJ: Hampton Press Inc, 1993. 237-56.

Race, Phil. Assessment: A Guide for Students. LTSN Generic Centre. Assessment Series. 2001.

Reckase, Mark D. "Portfolio Assessment: A Theoretical Estimate of Score Reliability." *Educational Measurement: Issues and Practice* (1995): 12-14, 31.

Reichert, A. "Teaching Students to Reflect: A Consideration of Program Structure." *Journal of Curriculum Studies* 22 (1990): 509-527.

Rossi, P. H., and H. E. Freeman. "Strategies for Impact Assessment." *Evaluation: A Systematic Approach*. London: Sage, 1993. 231.

Sadler, D. Royce. "Formative Assessment and the Design of Instructional Systems." *Instructional Systems* 18 (1989): 119-144.

Schon, D. *Educating the Reflective Practitioner: Towards a New Design for Teaching and Learning in the Professions*. San Francisco: Jossey Bass, 1987.

Schon, D. *The Reflective Practitioner: How Professionals Think in Action*. London: Basic Books, 1983.

Shulman, L. Portfolios in Historical Perspective. Presentation at the Portfolios in Teaching and Teacher Education Conference, Cambridge, MA 1994.

Simon, Marielle, and Renee Forgette-Giroux. "Impact of a Content Selection Framework on Portfolio Assessment at the Classroom Level." *Assessment in Education* 7 (2000): 83-101.

Simons, H. *Getting to Know Schools in a Democracy: The Politics and Process of Evaluation*. Lewes: Falmer Press, 1987.

Simons, Helen. "The Paradox of Case Study." *Cambridge Journal of Education* 26 (1996): 225-240.

Smagorinsky, Peter, ed. *Speaking About Writing: Reflections in Research Methodology*. Vol. 8. London: Sage, 1994.

Smith, Kari, and Harm Tillema. "Evaluating Portfolio Use as a Learning Tool for Professionals." *Scandinavian Journal of Educational Research* 42 (1998): 193-205.

Smith, Kari, and Harm Tillema. "Long-term Influences of Portfolios on Professional Development." *Scandinavian Journal of Educational Research* 45 (2001): 183-203.

Smith, M. L. "Put to the Test: The Effects of External Testing on Teachers." *Educational Researcher* 20 (1991): 8-11.

Smith, William L. "Assessing the Reliability and Adequacy of Using Holistic Scoring of Essays as a College Composition Placement Program Technique." *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Ed. M. M. Williamson and Brian Huot. Cresskill, NJ: Hampton, 1993. 142-205.

Snadden, David, and Mary Thomas. "The Use of Portfolio Learning in Medical Education." *Medical Teacher* 20 (1998): 192-199.

Snyder, Benson R. *The Hidden Curriculum*. Cambridge: MIT Press, 1970.

Stake, R. A. "An Approach to the Evaluation of Instructional Programs (program portrayal vs. analysis)." *Beyond the Numbers Game: A Reader in Educational Evaluation*. London: Macmillan, 1972. 161-62.

Stake, R. E., and D. Kerr. Rene Magritte, Constructivism, and the Researcher as Interpreter. Paper presented to the Annual Meeting of the American Educational Research Association, New Orleans, April, 1994.

Stecher, Brain. "The Local Benefits and Burdens of Large-Scale Portfolio Assessment." *Assessment in Education* 5 (1998): 335-351.

Supovitz, Jonathon, Andrew MacGowan, and Jean Slattery. "Assessing Agreement: An Examination of the Interrater Reliability of Portfolio Assessment in Rochester, New York." *Educational Assessment* 4 (1997): 237-259.

Taylor, Catherine S. "Using Portfolios to Teach Teachers about Assessment: How to Survive." *Educational Assessment* 4 (1997): 123-147.

Tierney, R. J., M. A. Carter, and L. E. Desai. *Portfolio Assessment in the Reading-Writing Classroom*. Norwood, MA: Christopher Gordon, 1991.

Tom, A. R. *Redesigning Teacher Education*. New York: State University of New York, 1997.

Underwood, Terry. "The Consequences of Portfolio Assessment: A Case Study." *Educational Assessment* 5 (1998): 147-194.

Valencia, Sheila, and Kathryn H. Au. "Portfolios Across Educational Contexts: Issues of Evaluation, Teacher Development and System Validity." *Educational Assessment* (1997): 1-35.

Vaughan, C. "Holistic Assessment: What Goes on in the Rater's Mind?" *Assessing Second Language in Academic Contexts*. Ed. L. Hamp-Lyons. Norwood, NJ: Ablex, 1991. 111-25.

Vygotsky, L. *Mind in Society: The Development of Higher Psychological Processes*. Ed. Michael Cole; Vera John-Steiner; Sylvia Scribner; Ellen Souberman. Cambridge, MA: Harvard University Press, 1978.

Vygotsky, Lev S. *Thought and Language*. Cambridge: Massachusetts Institute of Technology, 1962.

Vygotsky, Lev. *Thought and Language*. Cambridge: Massachusetts Institute of Technology, 1962.

Wade, Rahima, and Donald Yarbrough. "Portfolios: A Tool for Reflective Thinking in Teacher Education." *Teaching and Teacher Education* 12 (1996): 63-79.

White, Edward. *Teaching and Assessing Writing*. San Francisco: Jossey Bass, 1994.

Whitford, B. L., G. Ruscoe, and L. Fickel. "Knitting it All Together: Collaborative Teacher Education in Southern Maine." *Studies of Excellence in Teacher Education: Preparation at the Graduate Level*. Ed. L. Darling-Hammond. New York, Washington, DC: National Commission on Teaching and America's Future and American Association of Colleges of Teacher Education, 1999.

Wiggins, G. P. *Assessing Student Performance: Exploring the Purpose and Limits of Testing*. San Francisco: Jossey-Bass, 1993.

Wiggins, G. "The Constant Danger of Sacrificing Validity to Reliability: Making Writing Assessment Serve Writers." *Assessing Writing* 1 (1994): 129-139.

Wolf, A. *Competence-Based Assessment*. Buckingham: Open University Press, 1995.

Wolf, D. P. "Assessment as an Episode of Learning." *Construction Versus Choice in Cognitive Measurement*. Ed. R. Bennet and W. Ward. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.

Wolf, D., et al. "To Use Their Minds Well: Investigating New Forms of Assessment." *Review of Educational Research* 17 (1991): 31-74.

Wolfe, E. W., and B. Feltovich. Learning How to Rate Essays: A Study of Scorer Cognition. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, 1994.

Wolfe, E. W., and M. Ranney. "Expertise in Essay Scoring." *Proceedings of ICLS 96*. Ed. D. C. Edelson and E. A. Domeshek. Charlottesville, VA: Association for the Advancement of Computing in Education, 1996. 545-50.

Wood, D. "How Children Think and Learn." *Thinking Voices: The Work of the National Oracy Project*. Ed. K. Normal. London: Hodder and Stoughton, 1988. 197-205.

Yorke, Mantz. "The Management of Assessment in Higher Education." *Assessment and Evaluation in Higher Education* 23 (1998): 101-116.

Yorke, Mantz. Assessment: A Guide for Senior Managers. LTSN Generic Centre. Assessment Series.

Zeichner, K., and R. Tabachnik. "Reflections on Reflective Teaching." *Issues and Practices in Inquiry-Oriented Teacher Education*. Ed. R. Tabachnik and K. Zeichner. London: Falmer, 1991. 1-21.

Zeichner, K., and D. Liston. "Teaching Students to Reflect." *Harvard Educational Review* 57 (1987): 23-28.

Zeichner, K. “Reflective Teaching and Field-based Experience. *Interchange,* 12 (1981): 1-22.

Glossary to Chapter Three

Formative assessment

This is assessment used during the learning process to move students' learning forward. Knight (2001) described various characteristics of formative assessment. This is used to "identify what learners need to do in order to improve their work" (Knight 2001, p.7). Traditional assessment recognises the importance of formative assessment and, although it perceives formative assessment as thoroughly "unreliable", this is perceived as being unimportant given the "low stakes" nature of formative assessment and opportunities instead that formative assessment offers for addressing the pressures described previously in, p.35-36 to address concerns about the appropriate learning contexts.

- "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" (Knight p.7).
- "Formative assessment purposes thrive on disclosure" (Knight 2001, p.7).
- "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" (Knight 2001, p.7). Many higher education tasks and outcomes are complex, fuzzy and ill-defined.
- "Good formative assessment therefore implies thinking about learning, teaching and assessment, not just about assessment" (Knight 2001, p.8).

Summative assessment

Traditional (positivist) assessment has traditionally been very interested in summative assessment which provides:

- "feedout"
- "high-stakes" assessment – "people are likely to do all they can to conceal their ignorance and suggest competence" (Knight p.3)

Summative assessment should be "accurate, objective and reliable" (Knight p.3).

Norm-referencing

This provides "data that allows us to rank achievements, comparing one student to another. Norm-referencing is comparative, telling us this student is better than another, similar to a third and not as good as a fourth" (Knight p.16)

Criterion-referencing

"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" (Knight p.17).

High-stakes assessment

Something important and external to the student, such as a degree certificate or professional qualification or permission to pass to the next year of study, depends on the results of the assessment.

Low-stakes assessment

Nothing important in terms of external rewards depends on the assessment. Probably the assessment is aimed at assisting the learning process. The assessment outcome may be important to the student in that it aids his/her development, but it does not have external implications.

Chapter 4
Conclusions

Lewis Elton and Brenda Johnston

This chapter can be very short. Any hope that we may have had when we started on the present investigation that we might be able to come up with some comparatively simple advice that would improve existing assessment purposes, structures and practices quickly disappeared. Chapter 2 has demonstrated that even the best of current practices are by and large not good practice, and this at a time when deficiencies in assessment are becoming a crucial managerial issue in universities. However, until management gives adequate time and resources for all academic teachers to engage in the kind of training and continuing professional development which the latter consider essential for every profession except their own, and academics are then prepared to engage in it, little of significance will change.

While we believe that only when training and continuing professional development - in all aspects of teaching, learning and assessing - becomes accepted as normal, the much more subtle challenges of chapter 3 can be fully faced, a start can and should be made at once to bring assessment into the 21st century. In demonstrating the overwhelmingly positivist basis of current approaches to assessment, which is in striking contrast not only to philosophical developments of the past half century, but even to current approaches to teaching and learning, we call into question the philosophical basis of virtually all current approaches to assessment. Admittedly, this basis is largely tacit – for most academics the fact that they have practised positivism in assessment all their lives resembles the attitude of M. Jourdain in Molière's 'Le Bourgois Gentilhomme', who "for more than forty years had been speaking prose without knowing it". This positivism has to become conscious, and then, hopefully, be discussed in connection with the relative positions of positivism and interpretivism in different contexts. One other matter that LE wants to draw attention to is the desirability of starting a serious discussion on the question whether the time may be ripe for degree classes to be abolished and replaced by profiles.

As if this proposed programme was not sufficiently difficult in itself, one other change is needed. The most positivist of all assessors today are not in academia; they are in the Department of Education and Skills, in the Higher Education Funding Councils and in the Quality Assurance Agency. The fact that they provide the money which pays the piper, does not mean that they should solely call the tune.

A boy wrote a poem

A boy wrote a poem,
It was from homework from class,
He wrote about cliff-tops,
And how the winds pass.
He just let it flow
from his head to his pen,
But his spelling was bad,
"C, do it again!"

A boy wrote a poem,
And thought of his mark.
And this time he checked it
And wrote in the dark.
He changed and corrected,
Gave it in the next day,
He got "B+ Good effort"
and threw it away.

When he wrote of the oceans,
They gave him an "E".
They gave him an "E"
for the tides of the sea,
"What does this mean?"
Said the boy to his work,

“Does it mean I’m just lazy,
Does it mean I’m a berk?”

When he wrote about sunrise,
They gave him an “A”.
They gave him an “A”
for the dawn of the day,
“What does this mean?”
Said the boy to his paper,
“Am I meant to be happy
To leap up, cut a caper?”

“What is this letter?
It is nothing to me!
It doesn’t mention the good bit
At the end of verse three”
And so thought the boy
But he couldn’t be sure
So he looked at his shoes
And the tiles on the floor.

Then before long
At a certain time,
They asked for the marks
of the efforts in rhyme.
To be written down
In a large orange book,
In symmetrical lines,
To be read at one look.

And he sat at the back,
At the back of the room,
Among the new novels,
The display work, the gloom.
And they asked him his marks,
And he read them ashamed,
When he got to the worst
“Tut tut” they exclaimed.

Oh, he thought.
But I don’t see what those marks have
Got to do with my work.

He still doesn’t understand.

Nicholas Chapman
(Aged 12)

Reprinted from the Times Educational Supplement, 16. 8. 1985.