BEYOND TESTING

TOWARDS A THEORY OF EDUCATIONAL ASSESSMENT

Caroline V. Gipps

ROUTLEDGE

Contents

Glossary

Assessment:	a wide range of methods for evaluating pupil performance and attainment including formal testing and examinations, practical and oral assessment, classroom based assessment carried out by teachers and portfolios.
Reliability:	the extent to which an assessment would produce the same, or similar, score on two occasions or if given by two assessors. This is the 'accuracy' with which an assessment measures the skill or attainment it is designed to measure.
Validity:	the extent to which an assessment measures what it purports to measure. If an assessment does not measure what it is designed it measure then its use is misleading.
Formative assessment:	takes place during the course of teaching and is used essentially to feed back into the teaching/learning process.
Summative assessment:	takes place at the end of a term or a course and is used to provide information about how much students have learned and how well a course has worked.
Ipsative assessment:	in which the pupil evaluates his/her performance against his/her previous performance.

Chapter 1

Assessment Paradigms

Introduction

Assessment is undergoing a paradigm shift, from psychometrics to a broader model of educational assessment, from a testing and examination culture to an assessment culture. There is a wider range of assessment in use now than there was twenty-five years ago: teacher assessment, standard tasks, coursework, records of achievement as well as practical and oral assessment, written examinations and standardized tests. There is criterion-referenced assessment, formative assessment and performance-based assessment, as well as norm-referenced testing. In addition, assessment has taken on a high profile and is required to achieve a wide range of purposes: it has to support teaching and learning, provide information about pupils, teachers and schools, act as a selection and certificating device, as an accountability procedure, and drive curriculum and teaching. These new forms and range of purposes for assessment mean that the major traditional model underpinning assessment theory, the psychometric model, is no longer adequate, hence the paradigm shift.

A paradigm is a set of interrelated concepts which provide the framework within which we see and understand a particular problem or activity. The paradigm within which we work determines what we look for, the way in which we construe what we observe, and how we solve emerging problems. A paradigm shift or 'scientific revolution' occurs when the old paradigm is unable to deal with an outstanding problem (Kuhn, 1970). This book is written as part of the attempt to reconceptualize assessment in education in the 1990s. There has been over the last decade an explosion of developments in assessment and a number of key actors have been reconceptualizing the issues. The aim of this book is to bring together much of this work to discuss and synthesize it in an attempt to further our understandings and practice in educational assessment: to develop the theory of educational assessment.

We need to develop a new way of thinking about assessment to deal with the issues that are emerging as assessment takes on this broader definition and purpose. For example, one outstanding problem which we have in assessment is how to reconceptualize traditional reliability (the 'accuracy' of a score) in terms of assuring quality, or warranting assessment-based conclusions, when the type of assessment being used is not designed according to psychometric principles and for which highly standardized procedures are not appropriate.

I use the term theory to refer to a vehicle for explanation and prediction, a framework that will allow us to understand, explain and predict. Theories, as devices for organizing and giving meaning to facts, are built up through the process of analytical work: abstract, conceptual analysis is the vehicle for isolating crucial dimensions and constituents. My aim is that through this analysis we will come to have a better understanding of the design, functioning, impact, as well as inappropriate uses, of assessment within the new paradigm.

It is important too, given the much wider and more significant role given to assessment, that these issues are made clear to a wider audience. This book is therefore aimed at all those who work in and around education and are interested in assessment: teachers and administrators, advisors, lecturers, policy makers and other educational researchers.[1]

In the chapters that follow I shall look at the technical issues, though at a conceptual rather than technical level, how assessment impacts on curriculum and teaching as well as its relationship with learning, criterion-referenced assessment, teacher assessment and performance assessment (and evaluate what they have to offer within the new paradigm), and questions of ethics and equity, before drawing together the analyses to put forward a framework for educational assessment. But first this chapter sets the scene by looking at purpose and fitness for purpose in assessment, the traditional psychometric paradigm and what we see as the new educational assessment paradigm.

Fitness for Purpose

I have already referred to reliability of assessment (by this we mean the extent to which an assessment would produce the same, or similar, score if it was given by two different assessors, or given a second time to the same pupil using the same assessor) which goes alongside validity (by this is meant the extent to which an assessment measures what it purports to measure) but there is more to testing and assessment than technical issues of reliability and validity. Assessments (which I use here to include tests,

examinations, practicals, coursework, teacher observations and assessment) come not only in a range of forms but with different purposes and underlying philosophies; these determine the range of appropriate use for an assessment. The first question to be asked then when considering the form of assessment to be used is 'what is the assessment for?' For example assessment to support learning, offering detailed feedback to the teacher and pupil, is necessarily different from assessment for monitoring or accountability purposes (for a start it is much more detailed). We must first ask the question *'assessment for what?'* and then design the assessment programme to fit.

I take the view that the prime purpose of assessment is professional: that is assessment to support the teaching/learning process. But, government, taxpayers and parents also want to know how the education system and individual schools are performing and they must have access to such information. A major, though not the only, element of this information is pupil performance as measured by tests and examinations. Assessment carried out for these purposes is likely to be more superficial since it needs to be relatively quick and manageable and needs to be more reliable than that to support learning. One can picture it as a form of survey (using postal questionnaires) as opposed to an in-depth study (using detailed interviews). Somewhere in between these two extremes of testing to support learning or for accountability purposes lies assessment for certification purposes, as with our public exams at 16 and 18: this assessment has to be both detailed (to provide comprehensive coverage) and reasonably reliable (so that we may have confidence that the results are comparable from one school to another and from one part of the country to another) though in other countries, for example Germany, this is not seen as an issue.

The problem that we have to confront is that tests designed for purposes other than to support learning—the huge quantities of multiple choice standardized tests in the USA, and the formal written exam in the UK—have had, we now realize, unwanted and negative effects on teaching and the curriculum. The stultifying effect of public exams on the secondary system in England has been pointed out by the HMI (1979 and 1988), and was a prime mover in the shift towards GCSE with its emphasis on a broader range of skills assessed, a lessening of emphasis on the timed exam and an opening up of the exam to a broader section of the age cohort. (All of this was brought in and supported by the same government which is now retrenching to a formal, exclusive, written exam, but that is another story). The limiting and damaging effect of standardized multiple-choice tests in the USA has also been well documented and analyzed in recent years (for example, Resnick and

Resnick, 1992). But assessment for monitoring and accountability purposes will not go away; on the contrary, a number of countries in the developing world are using assessment even more to gear up their education systems: in the USA, in New Zealand, in Australia as in Great Britain governments have linked economic growth with educational performance and are using assessment to help determine curriculum, to impose high 'standards' of performance and, in New Zealand and Britain, countries which have taken on board the New Right marketplace model, as a market signal to aid parental choice and competition between schools (Murphy, 1990; Willis, 1992a).

Mindful of the distorting effects of assessment for these purposes, the task assessment specialists must address is how best to design accountability assessment which will provide good quality information about pupils' performance without distorting good teaching (and therefore learning) practice. We must also explore other forms of assessment which can be used alongside accountability assessment to support learning, and criteria by which we can evaluate them. This is not to say that traditional standardized tests and examinations have no role to play in assessment policy, but that we need to design assessment programmes that will do what is required of them and have a positive impact on teaching and learning.

This brings us to the second question which should be asked, but almost never is: *what kind of learning do we wish to achieve?*' for we know now that different forms of assessment encourage, via their effect on teaching, different styles of learning. If we wish to foster higher order skills including application of knowledge, investigation, analyzing, reasoning and interpretation for *all our pupils,* not just the élite, then we need our assessment system to reflect that.

But a failure to articulate the relationship between learning and assessment has resulted 'in a mismatch between the high quality learning described in policy documents as desirable and the poor quality learning that seems likely to result from associated assessment procedures' (Willis, 1992b, p. 1).

We need to put on to the assessment agenda issues of learning style and depth. We must articulate the model of learning on which we are to base new developments in assessment over the next decade if we are to develop a sound model and one which will achieve the results we wish for it. After all, the original psychometrics was based on a theory of intelligence, while multiple choice standardized tests were based on a behaviourist model of learning: educational assessment for the next century must be based on our best current understanding of theories of learning.

In considering assessment paradigms I shall look first at the

traditional psychometric model, which is where testing in education began, and then look at what has come to be called educational assessment and how it differs from the psychometric model.

Psychometrics

The science of psychometrics developed from work on intelligence and intelligence testing. The underlying notion was that intelligence was innate and fixed in the way that other inherited characteristics such as skin colour are. Intelligence could therefore be measured (since it was observable like other characteristics) and on the basis of the outcome individuals could be assigned to streams, groups or schools which were appropriate to their intelligence (or 'ability' as it came to be seen). Thus the traditional psychometric testing model was essentially one of limitation: measuring attributes which are a property of the individual and which were thought to be fixed. This notion of limitation is seen now to be a major disadvantage of the psychometric approach. Assessment to support learning, by contrast, aims to help the individual to develop and further his/her learning: it is enabling rather than limiting. Another feature of psychometrics is the interpretation of scores in relation to norms: norm-referencing grades an individual's performance in relation to that of his/her peers, that is in terms of relative performance rather than their absolute performance. Norm-referenced tests are designed to produce familiar proportions of high, medium and low scorers. Since students cannot control the performance of other students they cannot control their own grades; this is now widely considered to be an unfair approach for looking at pupils' educational performance.

With the psychometric model comes an assumption of the primacy of technical issues, notably standardization, reliability and limited dimensionality. If individuals are to be compared with one another then we need to be certain that the test or assessment was carried out in the same way for all individuals, scored in the same way and the scores interpreted in the same way. Standardization is thus vital as is the technical reliability of the test within this model. These requirements can have a negative effect on the construct validity and curricular impact of the test since only some material and certain tasks are amenable to this type of testing.

Along with psychometric theory and its formulae and quantification comes an aura of objectivity; such testing is scientific and therefore the figures it produces must be accurate and meaningful. The measurements which individuals amass via such testing: IQ scores, reading ages, rankings etc. thus come to have a powerful labelling potential.

But the psychometric paradigm has two other problematic assumptions which have been articulated more recently (Berlak *et al.,* 1992; Goldstein, 1992 and 1993).

The first is the *assumption of universality,* which means that a test score has essentially the same meaning for all individuals; this implies that a score on a standardized reading test represents the individual's ability to read (the performance is extrapolated from the test to reading in the general sense) and that what this means is universally accepted and understood.

A key factor in this argument is the 'construct'; a construct is a term used in psychology to label underlying skills or attributes. A construct is an explanatory device, so-called because it is a theoretical construction about the nature of human behaviour. In test development the construct being assessed is defined before the test is developed: this is to make sure that the test assesses the attribute that it is supposed to, that it is 'valid'. In the case of reading a detailed definition of the construct 'reading' would include accuracy and fluency in reading both aloud and silently, comprehension of material, interest in reading etc. Thus a test which had high construct validity (i.e. which actually assesses reading adequately) should address each of these aspects of the skill. In fact, standardized tests of reading tend to assess only one aspect of the reading skill, for example, comprehension of simple sentences. This means that such a standardized reading test score does not represent the individual's ability to read in the widest sense, and therefore that the meaning of the score cannot be universally understood (since the user of the score would need to know which aspect of reading had been tested).

The second assumption is that of *unidimensionality* which relates to the conceptualization of constructs and impacts on the techniques used for analyzing test items. The assumption (within psychometric theory) is that the items in a test should be measuring a single underlying attribute. Thus when items are designed for a test they are first screened for obvious biases in terms of stereotypes either in the language or the pictures. The 'pilot' test is then given to a sample of students (which should be similar in characteristics to the intended sample). Item analysis is then carried out to get rid of items which are 'discrepant' i.e. items which do not correlate highly with the total score, because the test is meant to assess only one attribute. Items which have a high correlation with the total score are said to have high 'discrimination' while those which have low correlations are poor discriminators and are usually either dropped or modified. This approach comes from factor analysis techniques and the aim with a 'good' test would be to produce one which had only one underlying factor. This practice has two effects: first

it implies an artificial simplicity of measured constructs since many attributes are in fact multi-dimensional as in the example of reading given above. Second, if the original group of items chosen actually measures more than one attribute and only a few items relate to one of these attributes these few items will inevitably have low correlation with the final score and therefore be eliminated from the final test. Thus they will be excluded from the test because they are different from the rest, the majority, of the items. The result will be a test measuring a single attribute, but the interpretations made from the score to a broader conceptualization of the construct will be invalid (and the measured construct will be determined by the original choice of items which might have been balanced in the direction of the second attribute which would then become the main attribute).

Since many of the attributes or skills which we measure in tests are multi rather than unidimensional we can see that forcing tests into a unidimensional structure is illogical (Goldstein, 1993) based as it is on the unproved assumption of unidimensionality. Item response models of item analysis, including the Rasch model, are predicated on the factor analysis model assuming a single underlying factor and this is the basis of critiques of these models (see Goldstein, 1992; Goldstein and Wood, 1989).

Around the 1950s the benefits of the application of psychological measurement in educational settings producing tests such as intelligence tests (including group tests used in the 11+) aptitude tests and the like began to be questioned. This criticism of the psychometric approach had two main foci. First the notion of limitation and the belief that tests are measuring a property of the individual; its focus was, critics argued, on the degree of ineducability of the child which arises from defects in the child or his/her home and parents rather than considering problems in teaching, curriculum, etc. (Meredith, 1974, quoted in Wood, 1986; Walkerdine, 1984).

The second was that the key feature of reliability requires the standardization of administration and tasks as well as scoring. Tests based on psychometric theory have as a prime requirement measurement properties amenable to statistical analysis: reliability and norm-referencing are the prime concerns. This has profound implications for the style of task assessed, the limited ways in which tasks can be explained to pupils and the required non-interaction of the tester. As a result of having to meet these requirements, issues of validity and usefulness to teachers have sometimes been overridden or ignored.

Around the time of the publication of Bloom's *Taxonomy of*

Educational Objectives in the late 1950 educators began to articulate a need for assessment which was specifically for educational purposes and could be used in the cycle of planning, instruction, learning and evaluation. This was termed educational measurement.

Educational Measurement

Wood (1986) cites Glaser's 1963 paper on criterion-referenced testing as a watershed in the development of educational measurement i.e. the separation of educational assessment from classical psychometrics. Glaser's paper made the point that the emphasis on norm-referenced testing stemmed from the preoccupation of test theory with aptitude, selection and prediction. Wood maintains that every development in educational assessment since Glaser's criterion-referenced testing paper is based on the criterion-referenced model. As the chapter on criterion-referenced assessment will show, there are enormous problems in the development of this kind of assessment, that results from criterion-referenced assessment can also be used for norm-referenced type purposes, and indeed norms are often used to set and interpret criteria of performance. But nevertheless, the point is well made, that in order to move away from a *norm* referenced approach the only other reference we have come up with is that of criteria or standards, whether the result is described as criterion-referenced assessment, graded assessment, or standards-referenced assessment. There are different philosophies and techniques underlying these approaches but what they all have in common is that they do not interpret performance in relation to norms.

Educational measurement, by contrast with psychometrics, aims to devise tests which look at the individual as an individual rather than in relation to other individuals and to use measurement constructively to identify strengths and weaknesses individuals might have so as to aid their educational progress.

To find out 'How well' rather then 'How many' requires a quite different approach to test construction. Wood's definition of educational measurement therefore is that it:

1 deals with the individual's achievement relative to himself rather than to others;
2 seeks to test for competence rather than for intelligence;
3 takes place in relatively uncontrolled conditions and so does not produce 'well-behaved' data;
4 looks for 'best' rather than 'typical' performances;

5 is most effective when rules and regulations characteristic of standardized testing are relaxed;
6 embodies a constructive outlook on assessment where the aim is to help rather than sentence the individual.

and is happy to accept that this is 'thinking not of how things often are but rather of how they might or even ought to be…' (Wood, 1986, p. 194).

Where Wood uses the term competence (rather than intelligence) he is referring to the product of education, training or other experience rather than being an inborn or natural characteristic, as intelligence. We could more comfortably now use 'attainment' or 'achievement'. He argues that a powerful reason why educational measurement should not be based on psychometric theory is that the performances or traits being assessed have different properties: 'achievement data arise as a direct result of instruction and are therefore crucially affected by teaching and teachers' (p. 190). Aptitude and intelligence, by contrast, are traits which are unaffected by such factors, he claims. Achievement data is therefore 'dirty' compared with aptitude data and should not/cannot be analyzed using models which do not allow for some sort of teaching effect.

Looking for best rather than typical performance (the fourth principle on Wood's list) relates to Vygotsky's *zone of proximal development*. In educational assessment tester and pupil would collaborate to produce the best performance of which the pupil is capable, given help from an adult, rather than withholding such help to produce typical performance.

This also relates to the competence/performance distinction: competence refers to what a person can do under ideal circumstances, while performance refers to what is actually done under existing circumstances, competence thus includes the ability to access and utilize knowledge structures, as well as motivational, affective and cognitive factors that influence the response. 'Thus, a student's competence might not be revealed in either classroom performance or test performance because of personal or circumstantial factors that affect behaviour' (Messick, 1984). Elaborative procedures are therefore required to elicit competence; examination procedures tend to produce non-elaborated performance, i.e. they test at the lower rather than upper thresholds of performance (a profoundly non-Vygotskyian notion). This competence/performance distinction is a useful one to make in the consideration of educational assessment, but so that we do not get drawn into the question of whether we can infer competence from performance (i.e. the

deep ability from the surface performance) we should think instead in terms of *best* performance. Wood concludes his paper with a plea for teachers to see test and examination results as saying something about their teaching rather than just about the pupil; he cites their reluctance to do this as the reason why teachers make so little use of test results (Gipps and Goldstein, 1983). 'How do you persuade teachers to trust tests?' is Wood's parting question.

What is interesting is to see is how the agenda has changed in only ten years since Wood's seminal paper: a major development in educational assessment in England is now teachers' own classroom based assessment, while in the USA it is 'performance' or 'authentic' assessment in which (the latter at least) the teacher is centrally involved. In other words, the teacher has moved centre stage as an actor in assessment rather than being a simple administrator of 'better' tests devised elsewhere, the scenario when Wood was writing. Because of these developments educational measurement has now become called more generally educational assessment; this is largely because 'measurement' implies a precise quantification, which is not what the educational assessment paradigm is concerned with. I shall now look at some of the key authors who have elaborated and defined educational assessment.

Educational Assessment

Glaser (1990) makes the case that assessment must be used in support of learning rather than just to indicate current or past achievement. Glaser's own work in the area of novice/expert performance indicates that there are characteristics of learners which differentiate experts from novices across a range of domains. 'As competence in a domain grows, evidence of a knowledge base that is increasingly *coherent, principled, useful* and *goal-oriented* is displayed. Assessment can be designed to capture such evidence' (*ibid,* p. 477). 'Assessment should display to the learner models of performance that can be emulated and also indicate the assistance, experiences and forms of practise required by learners as they move towards more competent performance' (*ibid,* p. 480).

The sort of assessment that Glaser has in mind here are: portfolios of accomplishments; situations which elicit problem-solving behaviour which can be observed and analyzed; dynamic tests that assess responsiveness of students to various kinds of instruction; and 'scoring procedures for the procedures and products of reasoning'. In other words we need a much wider range of assessment strategies to assess a broader

body of cognitive aspects than mere subject-matter acquisition and retention (for a more detailed discussion of the nature of assessment to reflect deep learning, higher order thinking and meta-cognitive strategies, see chapter 2).

Glaser's point is that assessment must offer 'executable advice' to both students and teachers; knowledge must be assessed in terms of its constructive use for further action. 'Once mastered, the skills and knowledge of a domain should be viewed as enabling competencies for the future' *(ibid);* in other words the assessments must themselves be useful and must focus on the student's ability to use the knowledge and skills learnt.

Raven on the other hand (Berlak *et al.,* 1992) argues that we must develop assessments which assess performance in relation to valued goals, rather than separating cognitive, affective and conative factors (and indeed failing to assess the latter two). He also argues that we need approaches which assess them in a unified way, since people do not become competent in activities which they do not value. Raven's general argument, that we should move outside the cognitive, is to be welcomed and resonates with some of the ideas from cognitive science and learning theory in relation to the importance of metacognitive processes in performance.

Goldstein (1992) argues that we need to stop seeing testing as a static activity which has no effect on the pupil. On the contrary, the pupil is participating in a learning procedure, he argues, and his/her state will be altered at the end of it. For example, successfully completing early items in a test might boost confidence and result in a higher overall performance than failing, or being unable to complete, early items. Thus we should have a more interactive model of assessment which does not assume that an individual's ability to respond to items remains constant during the test. The more 'authentic' the assessment becomes, Goldstein argues, the more important it is to question the assumption that nothing happens to the student during the process of assessment.

'Authentic assessment' is a term used largely in the USA where the intention is to design assessment which moves away from the standardized, multiple-choice type test towards approaches where the assessment task closely matches the desired performance and takes place in an authentic, or classroom, context. Performance-based assessment, more commonly called performance assessment, aims to model the real learning activities that we wish pupils to engage with, for example, written communication skills and problem-solving activities, so that assessment does not distort instruction. Chapter 6 deals in detail with performance assessment, but briefly the intention in performance

assessment is to capture in the test task the same demands for critical thinking and knowledge integration as required by the desired criterion performance. The Standard Assessment Tasks outlined in the blueprint for the National Curriculum assessment programme in England and Wales (DES, 1988) are good examples of performance assessment. Performance assessments demand that the assessment tasks themselves are real examples of the skill or learning goals, rather than proxies. They support good teaching by not requiring teachers to move away from concepts, higher order skills, in depth projects etc to prepare for the tests. The focus is more likely to be on thinking to produce an answer than on eliminating wrong answers as in multiple choice tests. '…insights about how to develop and evaluate such tasks come not from the psychometric literature…but from research on learning in subject matter fields' (Shepard, 1991). However, when such tasks are required to support psychometric principles such as reliability and standardization, in order to be used in accountability settings, they fall short since that is not the purpose for which they have been designed.

The issue for performance assessment, as some see it, is how can tasks developed from, for example, diagnostic interviews be adapted for large scale administration and offer some level of confidence in comparability of results (which is necessary for accountability purposes). An alternative view is that we cannot force performance assessment into a psychometric model and that what we need is a range of approaches: more formal testing on a psychometric model for monitoring and accountability purposes and teacher-based approaches on an educational assessment model for assessment to support learning. This still leaves us with the question of whether assessment for certification and selection purposes can be more broadly conceived (as for example, the GCSE) to offer both beneficial impact on teaching *and* sufficient reliability for public credibility.

The dilemma that we face is that there are increased demands for testing at national level which must offer comparability, at the same time as our understanding of cognition and learning is telling us that we need assessment to map more directly on to the processes we wish to develop, including higher order skills, which makes achieving such comparability more difficult. Attempting to resolve this dilemma is part of the purpose of this book. There is no doubt we are faced with a paradigm clash, and the question is whether educational assessment can offer high quality assessments for a range of purposes.

In relation to our first question 'assessment for what?' Stiggins (1992) is one of those who take the view that assessment for accountability purposes and classroom-based assessment are so

It is an exceptionally thoughtful assessment of assessment and I am (along with anyone else who broods about education) much in your debt.' Jerome Bruner, personal communication with the author

When this award-winning book was originally published in 1994, a review in the TES said: 'Beyond Testing is a refreshingly honest look at the dilemmas facing those who are trying to make educational assessment more supportive of high-quality learning for all pupils and students … It contains powerful and practical messages for assessment developers, policy-makers, teachers and pupils. It exposes the very different agendas of those who wish to achieve greater system-wide accountability through educational assessment, and those who wish to use it to promote improvements in the quality of pupil learning.'

Routledge

978-0415689564
9 78 04 1 56 895 64