# 13
# Psychometric test theory and educational assessment

*Harvey Goldstein*

## Types of assessment

A basic distinction underlies what I shall have to say between assessment connected to learning and assessment separated from learning. In the former case, which I call *connected assessment*, there is a further distinction between assessment as part of learning and assessment as terminal evaluation of learning – a distinction roughly the same as summative and formative assessment. In the case of what I call *separate assessment*, its defining principle is its deliberate attempt to avoid connection with particular learning environments (e.g. curricula); perhaps the most celebrated example is the IQ test.

It is not my purpose to argue the merits or demerits of separated assessment, but rather to explore how the assumptions which historically have underpinned it have conditioned also our approaches to connected assessment. At one level my argument is about the details of the doing of assessment, and at another level it is about the role of a technological discourse in persuading people to accept a notion of objectivity based upon mathematical theory.

While my main concern is to explore the historical role of a particular kind of separated assessment, namely that emanating from psychometrics, other kinds are clearly possible, if not actually thriving. It is simply that because the psychometric model provides the most potent influence on our thinking it has become very difficult to separate that model from our whole understanding of what assessment is or can be. In supporting that statement I shall attempt to show why it is important to extend our view of assessment beyond this historically constrained framework. Much of what I have to say can be regarded as developing further Wood's (1986) lucid discussion on the theme of educational versus psychological assessment.



## The psychometric tradition

The development of so-called 'test score theory' as a branch of psychometrics has been concerned with proposing ever more sophisticated mathematical models for the behaviour of exam or test questions or items. This much is clear, though we still lack a complete historical account of this development (for useful contributions, see Gould, 1981; Sutherland, 1980; Mackenzie, 1981). Having its roots in cognitive psychology, test score theory is inevitably concerned with making inferences about mental processes and has chosen to do this by hypothesizing mental 'attributes' which are possessed by individuals, either in continuous or discrete form. Common examples are concepts of 'verbal reasoning ability' or 'IQ' itself. Moreover, it is clear that these labels are not mere conveniences attached to collections of items which are assembled to form a test, although we *could* regard them as such if we so wished. The mathematical models of test score theory in fact assume that such attributes really exist. Indeed, in a strong sense, these attributes *only* exist via the mathematical models, since they cannot be directly observed nor defined in other ways. In this sense, the term 'theory' is applied correctly, being understood to mean 'mathematical theory'.

In the next section, I shall outline the assumptions of test score theory and follow this by exploring how they have come to influence our basic understanding of assessment.

## Assumptions of test score theory

The 'classical' treatise on mental test score theory is still the book by Lord and Novick (1968), even though it is now 20 years old. In this book the authors give a detailed account of methods for constructing and analysing tests, and set out the mathematical and statistical assumptions of such procedures. They make an interesting distinction between 'weak' and 'strong' test score theory. The former is applied to the collection of so called 'item analysis' techniques which have been used by most test constructors for choosing items, calculating test reliabilities, etc. The term 'strong' is reserved for what has now become known as Item Response Theory (IRT; Lord, 1980), and which has been an area of considerable development in the last 20 years. It has been applied both to attributes which are supposed to have a continuous distribution and to those which assume only a small number of discrete 'states'. An example of the latter are so-called 'mastery' models, and the best known example of the former is the 'Rasch' model.

The older item analysis procedures (see, e.g. Anastasi, 1968) consist of a collection of indices designed to measure concepts such as reliability, discrimination, difficulty, etc. One virtue of IRT is that, unlike the older item analysis procedures, it is fully explicit about its mathematical assumptions. I will therefore address most of my remarks towards IRT. Nevertheless, what I have to say concerning IRT can be read as applying to all statistically based test score

theories. Thus, item analysis procedures are also based on an implicit mathematical model which shares all the basic features of IRT, and one of the most serious weaknesses of Lord and Novick is their failure to point out that the *essential* difference between the weak and strong models lies in the particular form of mathematical relationship which is assumed to exist between the observed responses to the test items and the underlying attribute values. Thus, depending on which mathematical relationship one decides to choose (e.g. arithmetic or logarithmic scale), one is led to somewhat different statistical procedures for estimating individual attribute scores or for judging the adequacy of a particular item. Such differences, however, are matters of technical detail and should not be allowed to obscure the fundamental unity of all test score theory. (A technical account of the differences which arise from choosing different forms of mathematical relationship can be found in Goldstein, 1980.)

## Dimensionality

The debate about the number of 'dimensions' of mental ability goes back to at least the 1920s involving Burt, Spearman and others in arguments about the uses of factor analysis. In simple terms, the number of dimensions of a test is equal to the number of distinct quantities belonging to each individual subject which are necessary in order to describe the observed relationships among the test item responses. (The original debate, it should be noted, concerned the number of dimensions underlying a set of test scores rather than item responses, but the ideas are the same.) Thus, a test is one-dimensional if the set of item responses (typically correct/incorrect) can be described in terms of a single ability value or, for example, mastery state for each individual, together with a set of 'parameter' values for each item (difficulties, discriminations, etc.), plus a random residual component representing unexplained variation.

Because they have underpinned virtually all of IRT to date, I will concentrate my discussion on unidimensional (continuous) models. More recent developments of multidimensional test theory models, often under the name of 'binary factor analysis' (Bartholomew, 1980; Bock and Aitkin, 1981), certainly extend the range of IRT but have had little empirical exposure to date. Nor do they represent a different *kind* of conceptual model, rather they provide a more realistic explication of the notion of mental attributes by allowing the possibility of several, rather than a single, dimensions.

If we study closely the idea of dimensionality, we immediately face a difficulty. In much of the mental test literature an (one-dimensional) attribute is assumed to be a property of an individual with respect to her responses to a set of items. Thus, a 'verbal ability' test might be regarded as such a set of items. Yet nothing in the mathematical theory actually allows us to make such an inference. The theory is concerned solely with the dimensionality of the test with respect to a particular 'population' of individuals. The specification of the population in fact is crucial. At one extreme it might consist of all possible individuals, but the choice of population must be made somehow. It is perfectly possible for a test to be unidimensional in one population and two-dimensional in another. Thus, one can imagine a case where a unidimensional test existed for each of several distinct populations, but where the average item responses differed among populations, i.e. the item difficulties changed. The population formed by combining these populations would then in general contain at least two apparent dimensions. For example, if we had one population where the order of difficulty of just two of the items in a test was the reverse of the order in a second population for all values of the underlying attribute, then the test could not be unidimensional in the combined population.



If we wished to infer something about attributes which could be regarded as something like mental characteristics possessed by individuals, then we should have to demonstrate, at the very least, that these attributes could be identified in every subpopulation of the population of interest, which is no mean task. In reality, therefore, the notion of dimensionality, and hence of mental attributes inferred from sets of item or question responses, is inextricably linked to a choice of population, and inferences about an individual's underlying attributes are very much concerned with the company she is expected to keep. This is not to deny that the explanation of dimensionality in particular populations is without interest, since it may well indicate the existence of relationships and perhaps provide insights into the reasons for such relationships. Whether such analyses tell us anything about the nature of mental processes is quite another matter.

In short, I am suggesting that mental test score theory is simply just a statistical device for summarizing (providing one or more weighted averages) a set of observed test item responses. It operates by making substantively arbitrary assumptions about the form of underlying mathematical relationships, a choice made typically on grounds of statistical convenience or mathematical elegance, with claims concerning dimensionality being relative rather than absolute ones. This point is important since the notion of individual attributes coinciding with measurable dimensions is a key one underlying common notions of 'standards', especially in connection with 'absolute' assessments. The widespread assumption that there really are attributes whose values remain only to be estimated has often seemed to be legitimated by the existence of a large body of mathematical theory predicated on this same assumption. Certainly, if these mental attributes really did exist then we should want something rather like IRT in order to describe them. It seems, however, that so close is the connection between test theory and assumptions about mental attributes, that our thinking has been conditioned to regard the existence of the former as requiring the existence of the latter rather than the reverse. Such legitimation may perhaps best be described in terms of shared cultural values. Given the dominant status of mathematical knowledge in Western culture, we need not be too surprised at the influence which psychometric ideas continue to have.

There is a further consequence of these views, which has echoes in much of the current assessment debate. The idea of mental attributes possessed by individuals implies that these attributes in a real sense exist independently of the

context in which they are observed. Thus, although the context may modify the expression of an attribute, this is essentially a practical problem for the test constructor. So, for example, a test or exam constructor might 'sample' a 'domain' by selecting items relating to a number of contexts and then carry out a suitable IRT or item-averaging procedure to predict the score on the underlying attribute. This view of context-free assessment is examined in more detail below.

## Standards

The topic of 'standards' in education is complex and I shall deal only briefly with one aspect; the attempt to relate different assessments, based on different instruments at different times, to a common measurement scale. The assessment might be a formal examination or one made by a teacher, but that will not affect my argument.

Clearly, if standards are to have any meaning then a common scale for expressing them is necessary, and for every relevant assessment it must be possible to convert it unambiguously on to this scale. Thus, for example, if standards in the new GCSE exam are to be compared over time, then the 'meaning' attached to each grade must remain the same. This implies that if a group of individuals (from a well-specified population) took two exams, we would expect on average that they should get the same grades on each exam. This, however, will only be possible if the two exams either both share a single dimension in the sense I have already described, or share the same proportionate mixture of several dimensions. Furthermore, in practice we require standards to be applicable across *different* groups of individuals (different exam boards or different years), and hence that all the relevant populations are essentially equivalent. If we add to this the diversity of methods of assessment and actual syllabuses, then such a requirement seems unlikely to be met. It needs hardly to be said that there is precious little existing evidence which bears upon this issue, nor is there any serious attempt on the part of those responsible for the GCSE to examine the issue. Similar comments apply to other current discussions of assessment, such as graded tests. This lack of awareness is surprising, given the extensive discussions in relation to GCE and CSE exams during the 1970s in England and Wales, and also an airing in some of the discussions during the formative years of the Assessment of Performance Unit (APU; Gipps and Goldstein, 1983).

The idea of test or exam 'equivalence' relies upon similar notions concerning one-dimensional attributes as does traditional test score theory, albeit without the same explicit models. Moreover, to judge by the experiences of the GCE boards in attempting to equivalence O-level exams, there is little empirical support for the theoretical possibililty. The idea that it is both desirable and possible to achieve equivalence – or maintain marking standards – nevertheless persists, just as it persists among those who continue to advocate IRT as a theoretical basis for assessment. Undoubtedly this persistence has considerable



practical convenience, and it would be interesting to study the reasons for it in more depth, but that lies beyond the scope of this chapter.

To the obvious question of how one might replace the IRT notion there are several responses. One, of course, is to reduce the importance of examinations and tests as selection instruments, so that the demand for equivalencing and standardizing is less urgent. Another response is to recognize the unreasonableness of the demand for equivalencing and adopt a different procedure. Thus, in the context of public exams in the UK, suppose that an exam is viewed as an index of achievement, a summary measure whose derivation is a matter for judgement and negotiation and where grading is with reference only to those who take that particular exam, with no attempt to equate grades across time or across exam boards. Then at least the implications can be made clear. While it may not be 'fair', in the sense that an achieved grade will depend upon which other candidates take the exam, it does at least share that property with most other aspects of social existence. As I have argued elsewhere (Goldstein, 1986), the openness of a well-understood procedure is both valuable and acceptable.

This discussion about standards in public exams lies within the theoretical framework of separated assessment rather than connected assessment, even though exams would appear to be connected in their intention through the medium of a syllabus. The reason is that in order to achieve equivalence between assessments which are linked to different learning environments (curricula), it has to be possible to separate an assessment from a *particular* environment, and this can only be done either by postulating effectively equivalent environments or environment-free assessments. Such a context-free view of assessment is precisely the inheritance which the standards debate has acquired from psychometrics. One may see this inheritance again in current debates about school records of achievement and profiles in the way in which terms like 'skills' and 'competencies' are proposed as context-free.

## Connected assessment

In contrast to the idea of, and the mathematical edifice concerned with, separated assessment, we have little of any kind of theory for a discussion of connected assessment. In fact, a very great deal of such measurement is carried out, much of it by teachers and students in the normal course of learning, some by the application of formal (e.g. diagnostic) tests, and some by researchers using more ethnographic type techniques. Its feature is its commitment to providing a specific evaluation of learning in a fairly well defined context, with little concern for external equivalence. Nevertheless, connected assessments may still be used to make particular comparisons across contexts where the contexts share 'common' features, which implies a judgement and hence a debate about the commonality of different features.

Connected assessment is a more general notion than that of 'formative' assessment though most of this is certainly connected; it can also include 'summative' assessment which is not necessarily intended to feed back to

learning. Because of its concern with a specific environment, however, it is generally unsuitable for selection purposes. To be suitable for selection, different measurements must be commonly scalable, and apart from particular common features, connected assessments are not designed to do this. It follows that any attempt to use connected assessments as selection instruments either leads to negotiated compromises which may contain political as well as theoretical justification, or in fact may destroy the connectedness of an assessment by forcing it to assume a separated form.

If this argument is accepted, then it raises serious questions about the role of teachers in separated assessment. It would be possible to encourage teachers to try to make separated assessments of their students on the basis of projects, etc., rather than by using external exams or tests, and certainly current GCSE proposals seem to make the assumption that this is feasible. If that is the case, and if teachers are still expected to make connected assessments in their teaching, then there arises some doubt about whether it is possible for teachers to carry out both types of assessment on the same students, and, further, whether the role of teachers may change if they are to shoulder much of the responsibility for separated assessment. Thus, for example, teachers may come to view student achievement more closely in accord with test theory tradition than formerly, so that notions of traits and underlying equivalences may come strongly to influence connected ways of evaluating learning.

## Conclusions

I have argued that there are two distinct kinds of assessment with different aims and procedures. Furthermore, they are so different that attempts to use one kind in circumstances that demand the other are inappropriate. I have suggested that public examinations in particular are an example of a basically connected assessment which has had the demands of a separated assessment imposed upon it, and most recently in a severe form. To remain connected, exams would in practice have to satisfy a number of requirements. First, there would have to be a single exam board, as in Scotland, and only one syllabus per subject would have to exist, so that explicit or implied equivalencing becomes unnecessary. Secondly, equivalencing across subjects and across time would have to be abandoned. The notion of fairness, which I suggest is the one which predominates in public exam assessment, would therefore be dependent on the content of the assessments and relative to the population of examinees. As such, it would become a matter for open debate, as I have suggested, where negotiated agreement rather than mathematical theory would be the prime mover. In this sense it is hardly different from most other forms of competition, and it is only curious why there should be any expectation that it should be otherwise. Of course, there will still be a need for interpreting exam results, but such interpretations can differ according to purpose and context, which actually suggests a more useful system than one which provides a single all-purpose assessment.



It is also clear that other forms of assessment such as graded tests and profiles face a similar dilemma to exams. If they are to be connected then their use in selection is open to some doubt, given typical current demands for selection devices, and if they are to be separated then this may undermine much of the reason for the current support which these developments are receiving from government and other funding sources. Above all, whether my thesis is accepted or not, there remains a real need for some serious thought about the directions in which assessment systems are being driven. We seem to be witnessing a great deal of activity developing new assessment instruments, but very little activity concerned with reflecting on where these developments are leading.

Finally, I do not argue here either for or against one or other kind of assessment in general. Rather, I am concerned to put each in its place and to understand when and where each may be used, and what assumptions underpin them. I am persuaded that existing psychometric test score theory is largely inappropriate for that part of educational assessment which claims to be separated, and is irrelevant to the large amount of educational assessment which is of the connected type. If we are to have theoretical bases for educational measurement, then alternatives to existing psychometric theory are needed.

## Acknowledgements

I am grateful to Patricia Broadfoot, Caroline Gipps, Alison Wolf and Bob Wood for their helpful comments on a draft on this article.

## References

Anastasi, A. (1968). *Psychological Testing*, 3rd edn. New York: Macmillan.

Bartholomew, D. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society*, **B42**, 293–321.

Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, **46**, 443–58.

Gipps, C. and Goldstein, H. (1983). *Monitoring Children*. London: Heinemann.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, **33**, 234–46.

Goldstein, H. (1986). Models for equating test scores and for studying the comparability of public exams. In D.N. Nuttall (ed.), *Assessing Educational Achievement*, pp. 168–84. Lewes: Falmer Press.

Gould, S.J. (1981). *The Mismeasure of Man*. New York: Norton.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Lord, F.M. and Novick, M. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison Wesley.

Mackenzie, D.A. (1981). *Statistics in Britain 1865–1930*. Edinburgh: Edinburgh University Press.

Sutherland, G. (1980). Measuring intelligence: English LEAs and mental testing, 1919–1939. In J.V. Smith and D. Hamilton (eds), *The Meritocratic Intellect: Studies in the History of Educational Research*, pp. 79–95. Aberdeen: Aberdeen University Press.

Wood, R. (1986). The agenda for educational measurement. In D.N. Nuttall (ed.), *Assessing Educational Achievement*, pp. 185–203. Lewes: Falmer Press.