The Mystification of Assessment

Harvey Goldstein

Since 1977 Harvey Goldstein has been Professor of Educational Evaluation at the University of London. Prior to that he worked at the National Statistical Service of the British National Institute for Child Development. Here he explores the so-called 'objectivity' of the Rasch model now favoured at the NFER in its work for the AFU.

A few years ago, educational assessment finally started to emerge from the dominance of IQ testing, and began to consider the need for carefully thought out educational as opposed to psychological motivations in devising assessment procedures. The subsequent development of notions such as criterion-referenced testing and process evaluation has demonstrated that a variety of tools can be devised, although many of these have been seen as providing qualitative rather than quantitative descriptions. For those who are interested in making comparisons among individuals and groups, however, some form of quantification of assessment seems necessary, and one of the principal tasks for educational statisticians, as I see it, is to provide acceptable quantitative tools for educational measurement. Nevertheless, since quantification without educational content is insubstantial, and since it is often all too easy to develop the former to a level which is technically sophisticated while losing the latter element, any new statistical technique needs to be evaluated carefully in terms of its basic premises and assumptions. The purpose of this article is to examine one such technique which has become known as 'objective measurement', on behalf of which some far reaching claims have been made.

A detailed critique of 'objective measurement' and the related 'item banking' methods in the context of national and local testing programmes has been made elsewhere (Goldstein and Blishen, 1977). Here, I shall explore the underlying philosophy of this approach and try to see what implications it might have in general for educational assessment. Before doing this, however, it will be useful to make a few brief comments on certain general critiques of educational measurement which seem to reflect a rather widespread misunderstanding of the term.

In a recent article, McAlister and Brown (1976), while making some apt comments about the use of psychometric techniques in measuring educational attainment, appear to equate attainment measurement itself with psychometrics, while they elsewhere refer to the 'assessment' of attainment as if this did not involve any 'measurement'. Of course, anyone is at liberty to suggest their own definition of the term, but it would be extremely unfortunate if a rejection of the psychometric approach to measurement were to lead automatically to the rejection of any measurement or quantification in education. Thus even the act of deciding whether or not a child has achieved a stateobjective can be viewed as a simple measurement process, and as soon as one is confronted with a number of such simple judgements, some logical ordering or summarization of them usually becomes necessary.

The relevant questions to ask are those concerned with the levels and types of measurement or quantifications which should be used to describe the defined attainments. For example, in criterion referenced testing we may wish to deal with a simple yes/no categorization, but this then leads on naturally to the calculation of percentages of individuals responding 'yes' and to comparisons of such percentages across groups of children. None of this activity is necessarily identical to the objectives embodied in the test used. Yet it does involve measurement and quantification, and hence the possibility of statistical analysis and summary. The statistical methodology itself has its inevitable connections with particular educational theories, even though its use has been often been to buttress such theories. That Seelnow's (1976) criticism of the 'psycho-statistical' approach seems to stem from his identification of particular techniques with certain approaches to educational measurement, and does not really stand up as a fundamental critique of the use of statistical methods as such.

Outstanding problems

To say that statistical methodology has a role in educational assessment, however, is not to deny that there remain considerable problems. Two of the principal difficulties which have always faced constructors of traditional educational tests, for example, have been those associated with the need to provide a balance between test items appropriate to children exposed to different curricula etc., and also with the problem of test items becoming outmoded with the passage of time. This latter difficulty, in particular, has thwarted attempts to make judgements about trends in educational attainment, since any apparent change might simply reflect the changing difficulty of the test rather than the achievement of individuals. Thus for example, any concern about so called 'falling standards' will have to face this issue, and it is in dealing with this particular issue that claims on behalf of 'objective measurement' have been advanced quite strongly.

Although nearly all the important developments in the methodology of objective measurement have come from Scandinavia and the United States, the British work in this area will be used as an illustration since its effects are more immediately apparent in this country. While several British educational researchers have been involved in objective measurement the most important centre of work lies inside the National Foundation for Educational Research (NFER) and most of the ideas behind the work of this group of NFER researchers are spelled out in the book by Alun
Wiltmott and Diana Powels. The Objective Interpretation of Test Performance (NFER, 1974). The following questions from Wiltmott and Powels give a fair summary of their underlying approach to objective measurement.

"An "objective" measurement is one of the types which is so familiar in the physical sciences, and to establish "objectivity" in a test, one requires first that the characteristics of the items in a test must be independent of the distribution of attainment in the group who are given the test. And, secondly, that the test should give estimates of attainment which are independent of the particular set of items which compose the test."

This seems to stress a devilish need to make education as much like physics as possible, and even embodies the rather curious notion (in the light of the Special Theory of Relativity) that the properties of a measuring instrument are invariant, and exist independently of the circumstances in which it is used. Even so, the authors are prepared to concede that such an ideal situation may not always exist and that some of the items in their tests may not conform to their ideal. Writing about their "model" (which I shall return to later) for an objective test, they say: "The situation in which items should fit the model, and not the model fit the items."

This is an extremely radical proposal, what is says to effect is that, given the particular assumptions embodied in the mathematical formula which relates a test item score to an individual's "underlying ability", any test item which does not conform to the general formula is simply discarded. This can only mean that educational reality is subject-test and can be interpreted in order to satisfy the model; rather than that the model is altered in order to better reflect any educational reality. In fact, Wiltmott and Powels make no serious attempt to consider how their mathematical model might be justified by any educational model, and after they have "thinned" their testing items in their tests they pay serious attention to whether what is left actually measures anything valuable.

Who are the misfits?

They even discuss what to do with those individuals who cannot appear to fit their model, and talk of:

Some form of discrimination on the part of those who knew he candidates well.

... other words, the "norm" consists of conforming to the model and those who do not are considered abnormal of needing diagnosis. Thus, a value judgement seems to be implicit and because the model tends to "fit" individuals who have broadly similar response patterns to test items, would not be too surprising if the "abnormal" individuals were left to belong to cultural minorities or so be those" following a novel curriculum. There is no discussion of the nature of such a possibility in Wiltmott and Powels, nor, for as I can tell, by any other proponents of objective assessment.

A "model" upon which the above utilities are based is known as the Rasch model, named after the Danish mathematician Georg Rasch. It is quite unnecessary to go into halcyon detail in order to describe the more important assumptions upon which this mathematical model rests can be found in Goldstein and Bischoff (1977).

They can simply be stated as follows. First of all it assumes there is a 'true' test of 'factor' which determines the "true" or individual responses correctly to an item in a test. For a mathematician test, for example, it would be assumed that something called 'mathematical ability' existed and could be described for an individual in terms of a single number. Every item in the test is likewise supposed to have a 'true' score or 'factor'. Secondly, the model assumes that the difficulty order of the items in a test remains the same for all individuals, whatever their back ground or their interests. Stated thus, the limitations of the Rasch model are fairly obvious, so that any use of it ought, at the very least, to be restrictive and exploratory.

APU adopts Rasch?

Nevertheless, advocates of objective measurement are not simply operating their time playing with their models inside research foundations, where it might be regarded as a pastime interesting and fairly harmlessly academic pursuit. Unfortunately, this is far from being the case. The NFER, for instance, is carrying out much of the monitoring work for the DES Assessment of Performance Unit (APU), and some of their more ardent researchers have been advancing the use of the Rasch model in the design and analysis of APU tests.

One particular proposal envisages a 'cookie' consisting of a very large number of items which will be calibrated against each other, and selections made according to a strict criterion. Resulting tests, it is claimed, will be suitable for use by teachers without further modification, and with a ready-made calibration available so that results can be scored on a common 'objective' scale. In particular, the results from test designed for different curricula can be. Compared with each other, so that all individuals can be ranked on a single scale (excluding presumably those who do not belong to these curricula), it is also highly questionable as to whether such a goal is even desirable.

Nevertheless, the possibility of such a absolute measurement scale has a certain attraction and it is unlikely to be abandoned easily, despite any lack of educational relevance. A second strong reason for a resistance to action in this arises from its claim that it provides a method of setting comparisons to 'sing as all the items used are selected from the same latent base with the "true" items being calibrated against "false" notes. This, however, is also impossible, even within the assumptions of the Rasch model, as the following simple argument shows.

Illogical claims refuted

If we suppose that each of the items in the bank has a prescribed difficulty value, then it is strictly meaningless within the context of a Rasch model to speak of one item as being more applicable to one point in time rather than to another. The only meaning which can be attached to such a statement must be in terms of difficulty values. For example, suppose there are two items, one of which is most applicable in 1975 than 1980 and the other of which is more applicable in 1980 than 1975. Then the two items will have different relative difficulties in these two years, and, indeed, their relative difficulties might become reversed between 1975 and 1980. Hence, by definition, they cannot belong to a single common Rasch scale extending over this five-year period. Nor will it be possible to "calibrate" their difficulties with other items whose difficulties, for the sake of argument, are assumed to remain constant. Thus as item
Dismissing simplicity

I have argued that the objective measurement movement in education is misguided and often over-simplistic solutions to complex problems. It is, however, the very simplicity of these solutions which is extremely seductive, and this is compounded by the jargon which surrounds the methodology and which has been an art of deflecting and precluding the one thing which is not revealed clearly by those who advocate the methodology, however, is just what the mathematics actually implies in educational terms. Certainly one searches in vain for such statements in Wilborn and Powel. It is in this way that the advocacy of objective measurement tends to lead to mystification, it takes the discussion of the curriculum and its evaluation out of the major educational forum, and essentially hands it over to technicians who can manipulate the mathematical equations. In the final analysis, even more than the use of objective measurement itself, it is this mystification which serves to pose the most serious threat to the integrity of education evaluations and assessment.

Acknowledgements

I would like to thank Timo Tordoff, Steve Babbington and Bob Foul for their helpful comments on an early draft of the article.

References


A Subscription to Forum

To: Forum, No 1 New Road, Linbyfield.

Please send the name of Forum, starting with Vol. No. 1 and check for £5.00.

Please send a Bankers Order or Form to the address below.

NAME:

ADDRESS:

* Addressee as appropriate.