FORUM, Sept. 1979

59

The Mystification of Assessment

Harvey Goldstein
Since 1977 Harvey Goldstein has been Professor of Statistical Methods at London University Institute of Education. Prior to that he worked at the National Children's Bureau and was principally responsible for the statistical analysis of the British National Child Development Study. Here he exposes the so-called 'objectivity' of the Rasch model now favoured at the NFER in its work for the APU.

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 insufficient, and since it is often all too easy to develop the former to a level which is technically sophisticated while lacking 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 Blinkhorn, 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 this term.

In a recent article, McIntyre and Brown (1978), 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 a 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 stated objective 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 summarisation 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 categorisation, 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 inimical 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 no inevitable connection with particular educational theories, even though its use has often been to buttress such theories. Thus Stenhouse's (1978) 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 outdated 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 be reflecting 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 Reasearch (NFER) and most of the ideas behind the work of this group of NFER researchers are spelled out in the book by Alan

Willmott and Diana Fowles, The Objective Interpretation of Test Performance (NFER 1974).

The following quotations from Willmott and Fowles give a fair summary of their underlying approach to objective measurement.

'An "objective" measurement is one of the type which is so familiar in the physical sciences.'

and to establish 'objectivity in a test,

'. . . requires first that the characteristics of the items in a test must somehow be made 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 comprise the test.'

This seems to express a deeply felt 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 criterion is that items should fit the model, and not that the model should fit the items.'

This is an extremely radical proposal. What it says in 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 pattern is simply discarded. This can only mean that educational reality is subservient and can be deformed in order to satisfy the 'model' rather than that the 'model' is revised in order to better describe any education reality. In fact, Willmott and Fowles make no serious attempt to consider how their mathematical model might be justified by any education model, and after they have discarded the 'non-fitting' items in their tests they pay scant attention to whether what is left actually measures anything sensible.

Who are the misfits?

They even discuss what to do with those individuals who 'o not happen to fit their model, and talk of;

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

1 other words, the 'norm' consists of conforming to the nodel' and those who do not are considered abnormal id needing diagnosis. Thus, a value judgement seems to e implied and because the model tends to 'fit' individuals ho have broadly similar response patterns to test items, would not be too surprising if the 'abnormal' individuals rned out to belong to cultural minorities or to be those llowing a novel curriculum. There is no discussion of the ngers of such a possibility in Willmott and Fowles, nor, far as I can tell, by any other proponents of objective asurement.

e 'model' upon which the above claims are based is own as the Rasch Model, named after the Danish mathetician Georg Rasch. It is quite unnecessary to go into hnical detail in order to describe the more important mptions upon which this mathematical model rests se can be found in Goldstein and Blinkhorn, 1977). y can simply be stated as follows. First of all it assumes there is a single 'trait' or 'factor' which determines the nce that an individual responds correctly to an item in a test. For a mathematics 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 single value known as its 'difficulty'. Secondly, the model assumes that the difficulty order of the items in a test remains the same *for all individuals*, whatever their backgrounds, their exposures to different curricula etc. Stated thus, the limitations of the Rasch model are fairly obvious, so that any use of it ought, at the very least, to be tentative and exploratory.

APU adopts Rasch ?

Nevertheless, advocates of objective measurement are not simply spending their time playing with their models inside research foundations, where it might be regarded as a possibly interesting and fairly harmless 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 senior researchers have been advocating the use of the Rasch model in the design and analysis of APU tests.

One particular proposal envisages a 'bank' containing a very large number of items which will be 'calibrated' against each other, and selections made according to a user's specifications. Resulting tests, it is claimed, will be suitable for use by testers 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 tests 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 fit!). Not only is such a possibility actually unattainable, it is also highly questionable as to whether such a goal is even desirable.

Nevertheless, the possibility of such an absolute measurement scale has a certain attraction and it is unlikely to be abandoned easily, despite any lack of educational relevance. A further strong reason for a reluctance to abandon it arises from its claim that it provides a method of making comparisons so long as all the items used are selected from the same item bank with 'new' items being calibrated against 'old' ones. This, however, is also impossible, even within the assumptions of the Rasch model, as the following simple argument shows.

over time/

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 more applicable in 1975 than 1980 and the other of which is more applicable in 1980 than in 1975. Then the two items will have different relative difficulties in these two years, and indeed their relative difficulty 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 via other items whose difficulties, for the sake of argument, are assumed to remain constant. Thus an item

bank based on the Rasch model and designed so that outdated items can be replaced is a self contradictory concept. A similar logic applies to claims that an item bank can be constructed to suit different curricula. Thus, despite its claims the methodology of objective measurement contributes nothing to the resolution of the difficulties facing test constructors, which were mentioned in the introduction.

Disarming simplicity

I have argued that the objective measurement movement in education is misguided and offers over-simple 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 about it an air of desirability and promise. 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 Willmott and Fowles. 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 main 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 seems to pose the most serious threat to the enterprise of education evaluation and assessment.

Acknowledgements
I would like to thank Tessa Backstone, Steve Blinkhorn and Bob Wood for their helpful comments on an early draft of this article.

References

1 Goldstein, H. & Blinkhorn, S. (1977) Monitoring Eucational Standards – An Inappropriate Model; Bull. Br. Psychol. Soc. 30, 309-311

2 McIntyre, D. & Brown S. (1978) The Conceptualisation of Attainment; Br. Educ. Res.J. 4, 41-50

3 Stenhouse, L. (1978) Case study and Case records: towards a contemporary history of education; Br. Educ. Res.J. 4, 21-40

4 Willmott, A.S. & Fowles, D.F. (1974) The Objective Interpretation of Test Performance (NFER)

Accountability – a contagious disease?

Joan Shapiro
Dr Joan P Shapiro is a Lecturer and Supervisor of Teacher Education at the University of Pennsylvania's Graduate School of Education in the USA. She wrote this article while on a year's leave of absence (1978/9) as an Honorary Research Associate in the Curriculum Studies Department at the University of London Institute of Education.

This article is an attempt to provide British educators with a synthesis of some major developments in the Accountability movement in education, focusing primarily on the US experience. It has been written because there appears to be a need to understand a trend, which is now spreading in the UK. Since the Accountability movement in the US is advanced by five to ten years on its British counterpart, it is hoped that an understanding of the American experience may provide some insights into the present UK situation.

In both the US and the UK, the term Accountability has been interpreted in diverse ways. The varied definitions range all the way from narrow monetary concerns to broad political connotations. However, irrespective of the definition utilized, I believe that Accountability is an understandable outcome of an unstable social and financial decade. During the 1970s, inherent mistrust of most institutions have led politicians, ratepayers, and the media to assume that they have the right to hold the school accountable not only for its spending but also for certain aspects of the educational process.

The punitive tone and the negative forms associated with Accountability in the US have led me to liken this trend to a disease. If viewed as such, then the rapidity of the spread of this movement suggests the Accountability is contagious in nature. Furthermore, in light of the negative attitude exhibited by the general public towards education, it is clear that the ability to ameliorate this condition is not good. Unless controlled, it is my contention that the Accountability movement will have a stultifying effect on education, inhibiting innovation and progress in this field for many years to come.

It may be argued that an analogy which compares Accountability to an infectious disease is a gross exaggeration. It may also be debated that the US is so different from Britain that reported abuses of Accountability in America could never happen in the UK. However, I would suggest that to ignore the causes, signs and symptons of Accountability in America would be very foolhardy. Indeed, even if only a few similarities between the two countries can be identified, then I believe that an analysis of the American Accountability movement merits some serious consideration.

Thus, in this paper, I shall discuss causes and effects of Accountability in the US; and I shall draw the reader's attention to similar developments already occurring in the UK. Finally, I will propose a treatment plan to help monitor and limit the growth of this movement.