British Educational Research Journal, Vol. 5, No. 2, 1979

*Consequences of Using the Rasch Model for Educational Assessment**

HARVEY GOLDSTEIN, *Institute of Education, University of London*

Constructors of educational tests have always contended with two major problems. The first arises from the varied aims and methods of teaching, each of which can make out a case for using its own particular test instruments designed to assess its own set of objectives. Thus, if we wish to test, say, arithmetic attainment of 8-year-olds we must decide what the test items are meant to be testing and we should choose a range of items which in some sense is 'fair' to all children. For example, if rote arithmetic items were chosen and the test then administered to children who had not been taught arithmetic in this way, the test could be said to 'discriminate' against these children.

The second problem, which really stems from the first, is the relative impermanence of all educational tests. A test which might be acceptable and appropriate at one time will eventually become dated, either because it becomes too familiar and children are 'taught the test', or because the individual items in the test themselves become outdated and inappropriate. One of the best known examples of this is in the study of trends in reading standards in England and Wales following the Second World War (Start & Wells, 1972). This study, using two tests to cover the whole period, found an apparent increase in reading standards up to the mid 1960s followed by no change in average test scores. This result, however, might not have been due to any real changes in reading standards, but rather to a decreasing relevance of the tests over time, owing to changing language use, teaching methods, curricula etc. (Burke & Lewis, 1975). Thus, children may have achieved lower test scores than expected because the test had become relatively harder, rather than because their attainments had fallen.

This state of affairs can be contrasted with that in the biomedical sciences where measuring instruments do not suffer these problems, at least not to the same degree. For example, a child's height is always measured by effectively the same well-defined instruments at all ages and times. Of course, there may be considerable debate over the usefulness of using such an instrument in the first place but, once that issue has

* This paper was originally given to a Schools Council/University of London Institute of Education Seminar on 24 October 1978.

This paper deals with a very controversial topic. Other contributions to the debate would be welcomed (Ed.).

been resolved, we tend to find a general agreement about its continued use in a variety of circumstances. This seemingly desirable situation has stimulated a number of those concerned with the construction and use of mental tests to find ways of approaching it, and the controversy over the use of the Rasch model is perhaps best understood in this context, as will be elaborated below.

In the first of the following sections there is an outline of so-called 'latent trait models' of which the Rasch Model is but one special case. This will be followed by a more detailed explanation of the Rasch Model itself, its assumptions and implications. Following this there is a discussion of procedures for assessing how well the Rasch Model works in any given context.

Although these sections are concerned with basically statistical models defined by mathematical equations, the models will be developed in a relatively non-technical fashion, using only simple algebra to convey the essential ideas. For those readers who wish to pursue some of the technicalities, references are given. Having described the model, the next section discusses some of the important educational consequences of using the Rasch Model and so-called 'item banks' based on it, in order to monitor educational achievement. Its relevance to the requirements of the Government's Assessment of Performance Unit (APU) is examined.

This article, perhaps inevitably, is essentially a critical one and I am aware that, in arguing against one particular approach, it may appear that I have nothing to replace it. In an attempt to meet this, there is a final brief section which suggests some alternatives, not only to the Rasch Model, but also to some of the more 'traditional' methods of constructing educational tests. Some of the proposals in that section owe much to the ideas of my friend and colleague, Bob Wood.

Latent Trait Models

Probably the best known and most widely used latent trait model is factor analysis and, because of its familiarity, a brief discussion of it will serve to introduce the essential properties of latent trait models.

The factor analysis model assumes that there are one or more underlying and unobservable factors or traits which characterise an individual and determine his or her observed responses. For example, we might suppose that a few basic traits characterise personality and could we but know what they were and measure them, we could determine how they affected an individual's behaviour. Although we cannot actually measure them, it is supposed that we can assemble measurable 'indicators' of them, which in some sense reflect their operation. In the well-known case of the general intelligence factor or IQ, for example, it is possible to measure individual performance on various tasks or items, the responses to which are assumed (or rather defined) to be determined by a set of underlying factors. Of these, general intelligence (g) is regarded as the most important. This model can be written down formally, but simply, as

$$\begin{aligned} x_1 &= a_1 g + b_1 h + \ldots \\ x_2 &= a_2 g + b_2 h + \ldots \end{aligned} \qquad (1)$$

where $x_1, x_2 \ldots$ represent the measured indicators, for example test items, $g, h \ldots$ represent the factors and $a_1, a_2, b_1, b_2 \ldots$ are the values of the 'coefficients' (or 'loadings') of the model. It is the problem of finding the values of these latter quantities towards which the theory of factor analysis is directed. (For simplicity the 'residual error terms' in the equation have been omitted.)

This model says that, for an individual, each measured indicator is simply a weighted sum of his factor values. For example, the weights for the first indicator (x_1) are the coefficients a_1, b_1 .... The model can also provide estimates of the factors (factor scores) for each individual in the analysis so that the values of '*g*', say, can be compared across individuals. In fact, the model is written just like a set of multiple regression equations, the main difference being that the factors *g*, *h* .... have unknown values. As it turns out this implies that in order to be able to estimate the values of the coefficients a_1, b_1 etc., we need to introduce further assumptions. These may take the form, for example, of giving some of the coefficients known *a priori* values (usually zero). Because of the general arbitrariness of any set of assumptions, and the fact that different sets of assumptions will generally lead to different results, factor analysis tends to be used most often as an exploratory technique which seeks to uncover the existence of possible factors, rather than to provide objective and definitive estimates of the coefficient values or individual factor values. It is worth noting that, in the special case where it is assumed that only one factor operates, the coefficients a_1, a_2 .... can all be estimated without introducing further assumptions. In practice it is rarely assumed that one and only one factor operates, but we shall return to this point below.

It is now a matter of history how some of the earlier advocates of factor analysis in the area of intelligence testing exaggerated its claims, especially with respect to 11+ selection tests. In large measure these exaggerations and the eventual reaction to them can be traced to a failure to recognise the limitations of the factor analysis model described above and to an attempt to attribute a strong objective reality, rather than an exploratory status, to the general intelligence factor.

Equations similar to (1) underly all latent trait models. In these, a set of indicators which may be, for example, total test scores or, as in the case of the Rasch Model, responses to individual items, is related to a set of unobservable factors or traits. The principal differences lie in the form in which factors are combined together and related to observed values, and in the way in which factors are assumed to be measurable, for example as continuous variables or as categories. Nevertheless, the above remarks about the various limitations of the factor analysis model apply in general to other latent trait models. The next section describes the Rasch Model in detail.

The Rasch Model and its Assumptions

Consider the simple case of a test which consists of, say, *n* items administered to *m* individual subjects. What we observe is the response of each subject to each item, which is assumed to be either a 'pass' or 'fail'. (The Rasch Model can be generalised to deal with more than two distinct types of response (Anderson, 1977), but this raises no essentially new issues relevant to present considerations.) We can set up a latent trait model relating the probability of a pass to two underlying factors or traits, which we label as the 'ability' of the subject and the 'difficulty' of the item. One simple such model can be written as

$$p_{ij}=b_i+c_j \tag{2}$$

where the index *i* takes the values 1, 2, ... *m* and refers to individuals, and the index *j* takes the values 1, 2, ... *n* and refers to items. Thus p_{12} is the probability that subject number 1 'passes' item 2 and equation (2) states that this probability is equal to the simple sum of the subject's ability and the item's difficulty. We note that the *x*s in

the factor analysis model equation (1) have become probabilities (or proportions of passes) and the factor coefficients have become the set of quantities denoted by the subject abilities and item difficulties. Those familiar with the analysis of variance will recognise (2) as an analysis of variance analogue of (1) when it is regarded as a regression type model.

Because the probability p_{ij} must always lie between 0 and 1 and for other reasons of statistical convenience, the following slightly modified format (2) is used in practice;

$$\log\left\{\frac{p_{ij}}{1-p_{ij}}\right\}=b_i+c_j. \tag{3}$$

Equation (3) is the one originally suggested by Georg Rasch (1960) and which has subsequently come to bear his name.

A number of points arise from equation (3). First, of course, the values of b_i and c_i are not directly observable and we have to use the data provided by the pattern of passes and failures in order to obtain estimates of them. Also, as in the case of factor analysis with just one factor, we can obtain such estimates without any further arbitrary assumptions. Nevertheless, several assumptions have already been made. The first, and most obvious one, is that only one term or quantity (b_i) is necessary to characterise an individual or, to put it another way, an individual's ability is 'unidimensional'. Likewise, every item has only one characteristic, its difficulty. Although various methods have been suggested for testing the assumption of 'unidimensionality' of difficulty, there has been little work on the problem of adequately testing the 'unidimensionality' of ability. Indeed, the usual methods of testing unidimensionality of difficulty are based essentially *on the assumption* of the unidimensionality of ability. Thus, given that a unidimensional ability is always assumed in the analysis of the model, in order to satisfy or 'fit' the model, the items used must relate only to one underlying dimension of ability. This immediately differentiates the Rasch model from the usual factor analysis model in which the dimensionality or number of factors is studied in the analysis itself. For example, in testing mathematics, if we believe that children's responses to geometrical items are determined by different mental processes to their responses to algebraic items, then we should not be using both types in a model such as (3). This unidimensionality assumption is therefore a stringent one since it must govern our choice of items at the outset. Thus, if we do not take care judiciously to select items, then the model may well not fit the data and, although it may be possible to force a fit and to achieve a single dimension by allowing the analysis to remove 'discrepant' items, there is no guarantee that any sensible interpretation can be placed on the end product. The second point to notice about (3) is that the relative difficulty of the items in a test is the same for all individuals. Thus, for example, the ratio of the probability of a pass on item 1 to the probability of a pass on item 2 depends only on the difficulty values of these two items. This is the second strong assumption made by the model. As we shall see in the next section, it is perfectly possible to have models which are unidimensional in ability which do not require this assumption. Hence, even if we were satisfied that a test tapped only one dimension of ability, in order to use the Rasch Model we would also require that, despite different experiences, learning sequences etc., the difficulty order of items was the same for every individual.

The third point concerns what is known as the 'local independence' assumption. This says that, for any individual, the response to an item is completely independent of his or her response to any other item. This again makes strong assumptions which

can be violated in a number of ways. For example, if the order of the items in a test affects their difficulties we would not expect local independence to be true.

Finally, it is worth noting that equation (3) seems to imply a symmetry between item difficulties and individual abilities so far as the model is concerned. In reality, however, this is not the case. If the local independence assumption is violated, it is because the responses to the items for any given individual will not be independent and not because, for any item, the responses of individuals fail to be independent, since the independence of individuals can be guaranteed by the selection of a proper random sample. Also, since interest ultimately focuses on the individual abilities, it is the dimensionality of ability which is of major interest rather than the dimensionality of the items.

For a technical discussion of some aspects of the issues raised in this section the reader is referred to Lord & Novick (1968).

Testing the Adequacy of the Rasch Model

Suppose we generalise equation (3) somewhat as follows;

$$\log\left\{\frac{p_{ij}}{1-p_{ij}}\right\}=a_j(b_i+c_i). \tag{4}$$

The term on the left side of the equation is a 'logistic' function of p_{ij} and both (4) and (3) are examples of logistic latent-trait models. Equation (4) introduces a further quantity, known generally as an 'interaction' term, characteristic of each item. The quantity a_j in this context is known as the 'discrimination' of the item and makes the model more flexible than (3), in particular it is no longer necessary to have a constant relative difficulty between items. Note, however, that (4) still assumes only one dimension for individual ability. Because of its greater flexibility we can expect this model to have a better chance than model (3) of fitting a set of test scores. Unfortunately, however, the price for greater flexibility is an increase in the technical problems of using it, and in particular it turns out to be impractical to construct a useful item bank based upon it (see next section). Nevertheless, in the analysis of a set of data we can carry out statistical tests to decide whether (4) fits the data significantly better than (3) (see Anderson, 1977). In fact, the so-called 'goodness of fit' tests which are usually applied to the Rasch Model are essentially testing just this, as indicated in the previous section.

In view of the other strong assumptions made by the Rasch model, namely local independence and unidimensionality of ability, it might be thought that serious effort would have been applied to developing tests for them. This seems not to have occurred and until it does, possibly with the help of computer simulations, we are left with an inadequate basis for fully testing the Rasch model.

In any large program to evaluate the Rasch model it will be important to compare the results of analysing samples with very different characteristics. We should, for example, apply the model to minority groups, disadvantaged children, etc., in order to see whether the same item difficulty values are found in every case. This again does not seem to have been undertaken to any extent in any field of education and, until it is, we will not be justified in claiming much generality for the model.

To conclude this section I would like to mention a philosophical issue which has arisen in the debate over the Rasch Model. Among others, Willmott & Fowles (1974) make the following statement, "The criterion (of the adequacy of the model) is that

items should fit the model, and not that the model should fit the items." Their emphasis, in other words, is on defining a test *in order that it fits the Rasch Model.* Thus the techniques they use for deciding upon the adequacy of the model are seen not as indicating that the model itself possibly might not apply, but rather as tools for rejecting those items of a test which behave differently from the core of items which do conform to the Rasch model. Hence the attainment which is being assessed is effectively defined by those items which happen to conform to the Rasch model. As was pointed out in the previous section, such a procedure holds no guarantee that the result will have a real life interpretation. The implications of such a philosophy, which chooses its test content primarily on statistical rather than educational grounds, is examined in the next section.

Implications of Using the Rasch Model for Educational Assessment

It will be useful to consider a particular educational area to illustrate how the Rasch model might work in practice.

Consider testing the arithmetic attainment of a national sample of 10-year-olds. First of all, it will be recognised that in devising a fair test we are not measuring solely an 'inherent' ability to do arithmetic, but rather the accumulated experience of each child. This will encompass the curricula he has been exposed to, the type of teaching he has received, his social background, etc., as well as anything one might think of as 'native ability' in the subject; a situation, as explained below, where we might expect the Rasch model to fail.

Suppose it were possible, however, for the sake of argument, to eliminate the comparative effects of differing curricula etc., for example by selecting a very homogeneous group of children. Would we now expect the Rasch model to provide a reasonable description of the responses to a suitable set of test items? Would we, for example, expect items to appear in the same order of difficulty for all children? The answer will depend on which items are used. It might be possible to select items which did fit the model, but it is by no means clear that this is the kind of test we could want to construct; unless, that is, we were seriously to adopt the philosophy referred to at the end of the last section. Indeed, Wood (1978) has shown how a series of purely random events can produce an excellent fit to the Rasch model, so that the existence of a well-fitting set of items does not necessarily mean that they are measuring anything meaningful. The items of the arithmetic test which fitted the model might reflect, for example, the same common response of the children to the education they had all been exposed to, whereas our real interest might be in those items which did *not* appear to have the same difficulty for all the children. Of course, this immediately raises the problem of how we should choose and then combine a set of items into a single measure of arithmetic attainment, and this point will be taken up in the final section. For now it is sufficient to point out that, even if we could construct a well-fitting Rasch model in a situation where we might expect this to be possible, there is still no necessary educational reason for preferring it over any other method of test construction.

The above arguments do not necessarily imply, of course, that the Rasch model may not be of considerable use outside the area of educational testing. Psychologists, for example, are often concerned to attempt to isolate unidimensional mental traits and the Rasch model may have an important exploratory role to play here. It is perhaps worth emphasising, however, that the preoccupations of the psychologist are not identical with those of the educationalist. As we pointed out earlier with factor

analysis, the adoption of models from the former discipline by the latter is not a straightforward procedure.

Once we leave the above example of the homogeneous group of children and consider the real world, the difficulties of applying the Rasch model multiply, and it seems *a priori* unlikely that, for example, a reasonable and fair set of items can be found which appear in the same difficulty order for all children. Indeed, the essence of many educational systems is the diversity of approaches whose actual aim is to create differential attainments among otherwise similar children, for example by way of the order of teaching or as a result of different pedagogical objectives. While this situation need not rule out the possibility of a single common assessment, it does seem to be at odds with the rationale underlying the Rasch model and, by extension, other unidimensional latent trait models.

Item Banks, the Assessment of Performance Unit and Trends over Time

If the Rasch model actually worked in education, with a large set of items, and applicable to all individuals in a particular subject area, then in principle we could determine the difficulty value of every item and store these in a 'bank' for future use. (For a description of such a proposed bank in mathematics see Purushothaman, 1976.) Thus a test could be constructed using any selection of the items suitably covering the range of ability expected in the subjects to whom it will be given. Since the difficulty values of the items are known, the 'difficulty' of the test is known, and an individual's ability is found simply by counting the number of correct responses or passes and then referring this score to a calibration table compiled for the particular test, which converts this score to an ability value on the common underlying scale.

Thus, a detailed system of tailor-made tests could be constructed, suitable for children following different curricula without the need for extensive standardisation. Moreover, there would also be absolute comparability over time since new, and more relevant, items could be calibrated and incorporated in the bank and out-of-date items dropped, with a common reference scale for all items remaining.

So much for the dream. The educational reality, as argued above and elsewhere (Goldstein & Blinkhorn, 1977), is altogether different and has to do with a world which is too rich and complex to be reduced, without distortion, to such a simple model. Even so, and despite this, there is an inherent flaw in the item bank concept which would make it unworkable in practice. 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 the Rasch model to speak of one item as being *more applicable* to one point in time rather than another. The only meaning which can be attached to such a statement must be in terms of difficulty values. For example, suppose we have two items, one of which is more applicable in 1975 than 1980 and the other which is more applicable in 1980 than in 1975. Then these two items will have different relative difficulties in the 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 the five-year period 1975–1980. Neither is it possible to 'calibrate' their difficulties via other items, whose difficulties, for the sake of argument, are assumed to remain constant. Thus an item bank which is designed so that out-of-date items can be replaced is a strictly non-Raschian concept. Similar logic applies to the so-called 'tailored testing' procedures mentioned above where it is claimed that items can be selected from an item bank to suit different curricula.

Since a major justification for the use of the Rasch model is the construction of item banks for just the above purposes, it is unclear what role this model might play once we accept that the purposes are unattainable.

Turning now to the Assessment of Performance Unit (APU) of the Department of Education and Science (DES), we find that one of the principal aims of its monitoring programme is to be able to compare educational attainments over time. It was originally proposed that this could be achieved by utilising an item bank based on the Rasch model (Kay, 1976). Not only can the Rasch model not fulfil this aim, there is no other simple procedure which will do so either. As pointed out in the first section, there is no absolute basis on which the comparisons can be made over time. This seems to be a basic fact of life and we really ought to accept it with a good grace and try to discover precisely what we *can* usefully say about changes over time. In the following final section I shall outline briefly an alternative direction in which we might profitably turn our attention.

Alternatives to Rasch

The above critique of latent trait models, and especially the Rasch model, starts from the assumption that the criteria which properly ought to determine the content of an educational test are primarily educational rather than statistical. Phrased in this way, perhaps few would disagree. In recent years, the idea of a predominantly educational motivation for test content has found one expression in the development of so-called criterion referenced tests. These abandon traditional test construction procedures in favour of tests composed of items, each of which is designed to assess an articulated educational aim. As they are normally employed, the tests simply report whether or not each student achieves these aims (or passes the items) and comparisons *between* students are avoided and population norms are not used. Nevertheless, it is perfectly possible to calculate population norms and to compare students using these tests. Indeed, not to do so is to impose an artificial and unnecessary restriction, and the most important aspect of criterion referenced testing is its emphasis on the assessment of specific educational objectives. This point will be elaborated below, following a brief comment on traditional test construction techniques.

Two of the more important concepts traditionally used in the construction of mental tests are that the items comprising the test should be as 'homogeneous' as possible, and that the test itself should have a high 'reliability'. The homogeneity criterion means that the items are all measuring 'the same thing', and here we meet again the idea of a single underlying dimension around which the test is constructed. The procedures for obtaining homogeneity, however, are less stringent than in the case of the Rasch model and the increased flexibility allows more scope for the operation of educational criteria. All the same, there is still the assumption that an educational test should be measuring only one underlying dimension. Wright (1977) puts the point in an extreme form; "if they do not (bear on a single common latent variable), then the set of items contains a mixture of variables and there is no simple, efficient or unique way to know their utility for measuring anything".

The reliability concept says that the test score should have a high stability. In a proper test–retest situation an individual's score should not change markedly in relation to the total variability in scores between all individuals in the population. That is, there should be small 'measurement error'. Such a requirement is common

to all kinds of measurements and is not one which in any way conflicts with educational content criteria.

If, now, we allow ourselves to move away from the doctrine of a single underlying trait, we can allow educational criteria properly to determine test content. We may decide, for example, that a test of arithmetic should consist of an equal number of 'new' mathematic and 'traditional' mathematic items, with equal weight being given to each item in computing the overall test score. There will almost certainly be more than one 'dimension' present in such a test. We might alternatively decide to give relatively greater weight to each of the 'new' mathematic items (or alternatively include more of them) on the grounds that these are thought to be more important educationally. Yet again, we might determine the relative item weightings in such a way that the total test score was the best predictor of, say, later achievement in a group of children. No one of these approaches could be said, *a priori*, universally to be superior to the others. Each will be appropriate in a different context and relevant to answering different types of questions. Incidentally the fact that an equal item weighting system happens to be the way in which a test score is computed if we use the Rasch model does not mean that we are, therefore, necessarily using the Rasch model just because we use equal item weightings in a test. Thus, Wright (1977) claims too much when he asserts that anyone who uses a system of equal weighting "is assuming that their items are in fact working in just the way modelled by Rasch, whether they realise and capitalise on that assumption or not".

If we decide to construct tests as described in the previous paragraph this does not mean that we can abandon other criteria for deciding what are good items to use. We would still wish to have a high reliability, and we should not want to include poorly discriminating items, for example items that all the subjects can easily pass. We also need to pilot any new test, and we certainly need good random samples on which to standardise it. The kind of test referred to in the previous paragraph can be likened to what is found in most public examinations which attempt to cover one or more syllabuses with a representative and 'fair' set of questions. It is also found in other areas, for example the Retail Price Index is based on a collection of separate indicators or items. While there may be disagreements on particular items or weightings, there is still a consensus that the index is a useful thing to have. Perhaps this is how we should come to think of educational attainment tests. Also, in addition to reporting test scores we might wish to report the detailed pass rates of items or small clusters of cognate items say, rather like the National Assessment of Educational Performance (NAEP) in the United States.

The above emphasis on the qualitative element in educational test construction seems to have several advantages. It focuses attention on where it belongs—the educational aims being assessed, and focuses any conflict of aims at a crucial point, namely where the measuring instrument itself is being constructed. Naturally, with such measuring instruments, test constructors and users will have to forgo the luxury of having absolute comparisons over time and instead will have to think, for example, of comparisons between groups of subjects at different times, using tests appropriate to each. For example, if interest lay in regional differences in arithmetic at two points in time, we might well wish to make comparisons on the basis of different tests, each one appropriate to the epoch referred to. If it was felt that, in 1980, an arithmetic test should attach less importance to mental arithmetic skills than one which was administered in 1950, then the more useful regional comparisons will be those which compare regional differences using the test given in 1950 with those differences found

using a separate but appropriate test in 1980. Indeed, in the case of the reading tests discussed earlier, an attempt to provide absolute comparisons has tended simply to result in inappropriate and out-of-date measuring instruments with a limited utility.

In conclusion, I am arguing for a shift of emphasis away from a concern with the development of mathematical models aiming at neat technical solutions, and towards a development of quantitative assessment techniques which are firmly rooted in qualitative educational objectives.

Summary

It is argued that the Rasch Model, and item banks based upon that model, constitute inappropriate tools for use in educational assessment. The paper discusses dangers to the educational system which could result should this model be used for routine educational monitoring.

The underlying assumptions of the model are discussed in a non-technical fashion, and it is shown how these are related to other statistical models such as factor analysis. The paper points out an inherent logical contradiction in the very concept of an item bank based upon the Rasch Model, a defect which appears to have been overlooked by the proponents of the model. Finally, the paper argues for the consideration of alternative paradigms in educational assessment which are based upon quantitative measurements, firmly rooted in qualitative educational objectives.

Acknowledgments

My grateful thanks are due to the following who read and commented on an early draft of this paper: Steve Blinkhorn, Anne Hawkins, Jan-Eric Gustafsson, Raimo Konttinen, Philip Levy and Bob Wood.

Correspondence: Department of Statistics, Institute of Education, University of London, Bedford Way, London WC1H 0AL.

REFERENCES

ANDERSON, E.B. (1977) Sufficient statistics and latent trait models, *Psychometrika*, 42, pp. 69–81.

BURKE, E. & LEWIS, D.G. (1975) Standards of reading: a critical review of some recent studies, *Educational Research*, 17, pp. 163–174.

GOLDSTEIN, H. & BLINKHORN, S. (1977) Monitoring educational standards—an inappropriate model, *Bulletin of the British Psychological Society*, 30, pp. 309–311.

KAY, B. (1976) Justified impatience, *The Times Educational Supplement*, 1.10.1976.

LORD, F.M. & NOVICK, M.R. (1968) *Statistical Theories of Mental Test Scores* (Reading, Massachusetts, Addison-Wesley).

PURUSHOTHAMAN, M. (1976) *Secondary Mathematics Item Bank* (Slough, NFER).

RASCH, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests* (Copenhagen, Danmarks Paedagogiske Institut).

START, K.B. & WELLS, B.K. (1972) *The Trend of Reading Standards* (Slough, NFER).

WILLMOTT, A.S. & FOWLES, D.F. (1974) *The Objective Interpretation of Test Performance* (Slough, NFER).

WOOD, R. (1978) Fitting the Rasch model—a heady tale, *British Journal of Mathematical and Statistical Psychology*, 31, pp. 27–32.

WRIGHT, B.D. (1977) Misunderstanding the Rasch model, *Journal of Educational Measurement*, 14, pp. 219–225.