British Educational Research Journal, Vol. 8, No. 2, 1982

The Rasch Model Still Does Not Fit

H. GOLDSTEIN, *University of London Institute of Education*
STEVE BLINKHORN, *Hatfield Polytechnic*

ABSTRACT *Certain criticisms of the use of the Rasch model have been queried in a paper by Bryce. We argue that these criticisms not only remain valid, but are reinforced by recent research on latent trait models. In particular, Bryce fails to present a clear conceptual argument and adopts a very inefficient procedure for testing empirically the fit of the Rasch model.*

A paper by Bryce (1981) takes issue with some of the reasoning in earlier papers (Goldstein & Blinkhorn, 1977, Goldstein, 1979) where it was argued that the Rasch model was generally unsuitable for use in educational assessment. He also presents empirical evidence in justification of the use of the Rasch model for test construction and analysis.

While we believe there is a certain amount of common ground between us, we also think that Bryce has partly misunderstood the position we have adopted, and ignored some serious weaknesses of the usual type of empirical analysis which he uses in his paper. We first deal with the conceptual framework.

The Rasch Model Assumptions

There seems to be no disagreement between us that the Rasch model is unsuitable with respect to large-scale item banking. This is of course the use of Rasch with which we were most concerned, and it has been pointed out (Goldstein, 1979) that the Rasch model ought properly to be regarded as a special kind of factor analysis model with possibly an important exploratory role. Again, we think there is no disagreement between us on the question of the need to test the assumptions of the Rasch (or indeed any other) model empirically. Beyond this, however, we seem to part company and we now take up the points made by Bryce.

Bryce accuses us of 'axiomatically' stating that the complexity of data must be greater than that allowed for by the Rasch model while also allowing that the validity of Rasch is open to empirical verification. What we actually suggested was that there were some obvious practical testing situations when the Rasch model assumptions were implausible—for example applied to a test containing 'new' and 'traditional' mathematics items administered to children subject to varying degrees of exposure to related types of curriculum. In one of our papers (Goldstein, 1979) there is an elaboration of some of the reasons for supposing that there would be rather few interesting situations where Rasch could be expected to fit well.

Bryce states: "the argument (about the usefulness of Rasch) hinges on evidence we have concerning the reasons for item discrimination patterns". Certainly, *if* a Rasch model holds and different discriminations are due to a few 'bad' items, then fitting the model will indicate which those items are. Our point, however, is a different one; namely that once one accepts that responses to different items are determined by different processes, one ought not to expect all the items to fit a simple unidimensional model like Rasch. Bryce states that Goldstein (1979) makes "very many assumptions when [he] doubts *a priori* that a reasonable and fair set of items can be found which appear in the same difficulty order for all children". This is quite correct, but in fact the assumptions made in that paper are quite clear, and they are associated with the particular assumption of the Rasch model which requires that all items appear in the same order of difficulty to all children whatever their exposure to different curricula etc. Unfortunately, Bryce nowhere attempts to provide a detailed argument on this point, simply suggesting that "perhaps the same difficulty order for items for all children *is* true with 'good' or 'effective' teaching?" In an educational system with a diversity of curricula and the freedom to teach in different ways and with different emphases, this appears as a remarkably naive statement.

Before turning to Bryce's empirical analysis we would like to comment on his use of the notions of unidimensionality and discrimination. In Bryce's example of test items, the responses to which may depend on reading ability and physics ability, he is essentially envisaging a two-dimensional model. Naturally, any method of analysis which assumes only a one-dimensional model will fail to distinguish properly the reasons for different discrimination patterns, but this applies as much to a Rasch analysis as to any other and we fail to understand how 'a Rasch analysis in such a situation will avoid the problem'.

This raises the more general point, made in a more recent paper not referred to by Bryce (Goldstein, 1980), namely that the Rasch model is only one of an infinite number of latent trait models available for fitting test item responses. That paper shows, with a simple example, how an alternative unidimensional model can 'fit' a set of data just as well as Rasch but give very different individual ability estimates. It also shows that a good fit to the unidimensional Rasch model can disguise the genuine existence to two separate dimensions. More generally, computing techniques have now become available to fit a new class of latent trait models (Bock & Aitkin, 1981) which allow us to adopt a fully exploratory approach to issues such as the degree of dimensionality, in just the same way that traditional factor analysis allows us to. Moreover, not only is this possible, it is also necessary in order to replace the traditional kinds of latent trait models, in particular the Rasch model, since these suffer from an inherent technical deficiency which results in biased estimates of individual ability (Goldstein, 1980). Thus, empirical analyses of data which fit only the Rasch model are extremely limited since they do not allow a sensitive or powerful exploration of the data. This point is well illustrated by Bryce's own example, to which we shall return.

Before discussing that example there is one further point which is worth making here since we have not seen it clearly stated elsewhere. Suppose that in reality we have a two-dimensional model, for example a mathematics test consisting of a mixture of geometry and algebra items where individuals vary both in terms of geometry 'ability' and algebra 'ability'. Suppose also that we wish to obtain a single 'mathematics' score for each individual. Clearly then we need to make a decision about the relative contribution of algebra and geometry items to this score. If we decide on equal contributions or weights then we might achieve this by putting the same numbers of algebra and geometry items in the test, and using the raw score. If we wished to weight the algebra items twice as much as the geometry ones then we might achieve this *either* by having twice as many algebra items in the test and using the raw score, *or* by having equal numbers and giving a score of 2 for every correct algebra item and a score of 1 for every correct geometry item. In fact we will always have to make a choice about relative weighting. Simply to use the raw score does not avoid the choice because in that case we still have to decide how many of each type to include. Furthermore, different relative weightings will in general produce different rankings of individuals, dependent on their relative algebra and geometry 'abilities'.

We see therefore that there is really nothing 'natural' about the use of the raw score in the situation, which is typical in education, where more than one dimension exists. The same argument applies to the Rasch model. Ability estimates from this are merely an 'order-preserving' transformation of the raw score so that whenever, in reality, more than one dimension is present there is an implied weighting of the items determined by the construction of the test. Thus, in the two-dimensional algebra and geometry example, we could, for example, readily 'fix' the test to have an overwhelming preponderance of one or other type of item, with a Rasch analysis leading to the 'non-fitting' of the minority type of item. The analysis would then be reflecting simply the implicit weighting we had chosen for the items rather than whether they were 'good' or 'bad'. This serves to underline the point that so long as we restrict ourselves to a one-dimensional model when the reality is multidimensional we will be unable in general to make sound inferences about the structure of the data.

Empirical Example

Bryce's data consist of two 40-item Scottish Certificate of Education papers in physics administered to 710 pupils in all.

The Rasch model was fitted to the test item responses using a criterion of fit suggested by Wright & Mead (1977) and showed that 38 and 37 out of 40 of the items in each year 'fitted' the model. The analysis goes on to show that items which teachers claimed were 'unsuitable' for their children fitted less well to the Rasch model than those items thought to be generally suitable by the teachers. This result, of course, is totally unsurprising and serves only to emphasise the point that factors such as unequal curriculum exposure argue against the use of Rasch.

Unfortunately, the analysis has several weaknesses. Firstly, as already pointed out, it does not explore the possibility of more than one dimension of ability. More importantly, however, the test of fit to the Rasch model is an extremely insensitive one. In a recent paper, Divgi (1981) has analysed several data sets comparing Wright's fit criterion, adopted by Bryce, with a more sensitive criterion, and found that the average percentage of non-fitting items went up from 16 to 69%! In fact Divgi showed that Wright's criterion has very low power and should not be used as a guide to the adequacy of the Rasch model. Thus Bryce's analysis, in common with those used by other proponents of Rasch, is quite inadequate. It uses a fit criterion which is poor at detecting non-fitting items, and it provides extremely little information of an exploratory kind, particularly concerning dimensionality.

In conclusion therefore, it is worth emphasising that we do seem to agree with Bryce on some important issues. Namely that the Rasch model is not appropriate for large-scale item banks and that empirical exploratory analysis rather than uncritical acceptance of assumptions is important. In criticising Bryce's paper we hope both to clear up misunderstandings about our conceptual stance and to point out the severe limitations of Bryce's (and most other people's) empirical analyses as a test of the adequacy of the Rasch model. What seems to be encouraging is the emergence of new models for carrying out valid exploratory analyses with the potential for allowing us to gain insights into data structures. As argued elsewhere (Goldstein, 1980) these new techniques still have their problems, but they do seem to be a major step forward in freeing test analysts from the rigidity of Rasch.

Acknowledgement

We are most grateful to Desmond Nuttall for his useful comments on an earlier draft of this article.

Correspondence: Harvey Goldstein, University of London Institute of Education, Bedford Way, London WC1H 0AL, England.

REFERENCES

BOCK, D. & AITKIN, M. (1982) Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm, *Psychometrika*, 46, pp. 443–460.

BRYCE, T. G. K. (1981) Rasch fitting, *British Educational Research Journal,* 7, pp. 137–153.

DIVGI, D. R. (1981) Does the Rasch model really work? Not if you look closely, paper presented at *Annual Meeting of National Council for Measurement in Education,* Los Angeles.

GOLDSTEIN, H. (1979) Consequences of using the Rasch model for educational assessment, *British Educational Research Journal,* 5, pp. 211–220.

GOLDSTEIN, H. (1980) Dimensionality, bias, independence and measurement scale problems in latent trait test score models, *British Journal of Mathematical and Statistical Psychology,* 33, pp. 234–246.

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

WRIGHT, B. D. & MEAD, R. J. (1977) BICAL: calibrating items and scales with the Rasch model, *Research Memorandum No. 23* (University of Chicago, Department of Education).