Unpublished

The irrelevance of IRT to the analysis of complex achievement data: A response to Beaton and Johnson.

by

Harvey Goldstein

Institute of Education
University of London

C:\MS\IEA\BEATON.DOC

September 6, 1990

1 Basics of IRT

Item response modelling, usually misleadingly referred to as item response *'theory'* seeks to summarise a set of responses to test items in terms of a single individual 'ability' together with estimates of item characteristics such as difficulty and discrimination. Recent mathematical and computing developments have allowed psychometricians to develop efficient methods for estimating these quantities, and current interest centres on multidimensional versions which are analogous to traditional factor analysis, but where the variables are dichotomous (correct, incorrect) responses rather than continuous measurements. The paper by Beaton and Johnson (1990) only deals with the case of a single underlying ability.

A recent summary of the essential characteristics of this class of models is given by Goldstein and Wood (1989) and here I will summarise the main points.

To clarify the essentials I will consider a simple item response model (IRM) where the individual ability is estimated by the total number of correct items, the item difficulty is estimated by the proportion of individuals who fail to get it right and the discrimination is estimated by the correlation between the item response and the ability. I shall refer to this as the basic item response model. Goldstein and Wood show that such estimates are not necessarily the most efficient ones for such a model but are satisfactory and are simple to understand. On the other hand, a major problem when discussing so called IRT models is that their advocates invariably appeal to the mathematical complexity of the models to justify their adoption. Beaton and Johnson (B&J) are no exception, referring to IRT 'technology', 'nonlinear functions' and the 'three parameter logistic model'. Somewhat surprisingly, they seem to be quite unaware that the basic item response model has all of the same essential characterics as their own 'three parameter logistic' model. Thus they dismiss the idea of using 'average percent correct' for a test without realising that the use of this can be derived directly from the use of the ability scores for individuals from the basic model. Thus, each individual has a score and the mean of these is essentially equivalent to the average percent correct. Furthermore, the distribution of scores can be calculated just as for the B&J model.

Even more surprisingly, they actually criticise the use of common items at two occasions to estimate trends when that is precisely what IRT does. B & J were recently involved in the so called NEAP reading anomaly which was based on such a procedure (Beaton and Zwick, 1990). To be fair, B&J are hardly alone in being unaware of the logical relationship between these different models. Nevertheless, this has led to a great deal of confusion and the remainder of this article will attempt to clarify the important issues.

2 Ability scales and model fit

B&J make various claims that their models 'fit' their data very well and claim to demonstrate that a single ability is an adequate summary of reading. McLean and Goldstein (1988) point out that the NAEP reading analyses were inadequate and that such analyses cannot justify any statement that a single ability (rather than many sub scales) is adequate. In any event, NAEP simply did not look very hard for more than a single ability. That again is a common characteristic of the proponents of IRT: they are too easily satisfied with oversimple summary descriptions. Interestingly, B&J themselves are critical of over simple summaries, but then fail to take their own reservations seriously.

In like manner B&J claim that a single scale of ability can be found which extends over a very wide age range. The supposed advantage of such a scale is that apparently comparable scores can be reported for all ages. It becomes rather like a measurement such as height or weight, so that we can talk about 'development' or 'growth' in similar terms. Unfortunately the justification for such scales falls apart when the logic upon which they are based is examined closely.

Consider two ages, say 7 and 11 years and a set of reading items at each age with 10% of items in common. Given appropriate samples we can determine the average difficulty of the items in both the 10% common set and also for each of the remaining items. We would expect the difficulty of the 10% to decrease from 7 to 11 years. Suppose the reduction in difficulty is x. If we subtract x from the difficulties of all the 7 year items we immediately make the average difficulties of the 10% equal and we can think of the remaining 90% as also being 'adjusted' to the give a single scale for all 7 and 11 year olds. The trouble is that this assumes that the relationship between the 10% and the 90% at 7 in some sense is the same as it is at 11 years. Generally speaking it is not, and can only be so if there is indeed a single underlying ability. That, however, is an assumtion which, as has already been pointed out, has yet to be verified. In fact, this whole procedure, usually known as 'vertical equating' is highly suspect. It is also inherently unreasonable when one considers the assumptions which are required about the nature of learning development in children. Moreover, even if the relationships were the same, this would constitute a necessary but still not sufficient justification for a single scale. To justify a single scale we would need a strong theoretical basis for supposing that there was such a single ability in reality and that the items chosen actually measured it. Mere data manipulation is not proof: nor does it in any way justify the epithet 'theory'.

3 Anchoring

The anchoring procedure supposes that students with a score of, say, 200, are alike. That assumes that a score of 200 is achieveable in only one way. In fact students can achieve this score by quite different patterns of responses. Even if there is a majority pattern, the departures are of interest, and to ignore these is to present a distorted and inadequate picture. It may be easy to explain anchor points to the world at large, but it misleading to pretend that they are good summaries and to fail to explain their shortcomings.

4 Context

It is important to decide how we aggregate separate items into clusters to form meaningful sub scores, and the global aggregation procedures of NAEP are misleading and wasteful of useful information. We need to be guided in such an activity by substantive educational criteria and not by atheoretical mathematical models. Once the actual choice of items for a cluster is made, the precise method of combining the responses to form a cluster or subscale score is not critical. Thus, there will often be little difference between the basic model I have described and the IRT model of B&J (Goldstein and Wood, 1989). Since the former is simpler to construct and easier to understand, there is a strong case for preferring it to the latter. The key issue is how to select the items, since this will determine the meaning to be attached to the scale. The most negative aspect of the B&J paper is that it diverts attention from this issue by overemphasising the importance of a particular mathematical model.

Once we have decided how to summarise item responses, we will wish to study the factors which are related to performance. We need to know how the environment of the school, its organisation, curriculum and teachers, influence the achievements of its children. To do this job properly we need to take careful measurements of such factors and to carry out sensitive multilevel analyses of the data. No amount of IRT fitting will avoid the hard work associated with this activity, in particular the difficulty of interpreting its results. Yet B&J have nothing to say about this. They give the impression that IRT, once carried out, solves all the major problems. It doesn't, and any pretence that it does is dangerous nonsense.

At all stages of data summary and data analysis we should be guided by what we know of educational theory and existing research. Those who are keen to adopt IRT distract us from this task with high sounding phrases about 'communicating information about the status of education' which promise much but mean little. It is a model which the IEA would be wise to reject.

5 references

Beaton, A.E. and Johnson, E.G. (1990). IRT as a way of improving the usefulness of complex data. Paper delivered to the annual meeting of the American Educational Research Association meeting, Boston, April 1990.

Goldstein, H. and Wood, R. (1989). Five decades of item response modelling. Brit. J. Maths. Stats. Psychol., 42, 139-67.

Mclean, L.D. and Goldstein, H. (1988). The U.S. national assessments in reading: Reading too much into the findings. Phi Delta Kappan, January 1988, 369-372.

Beaton, A.E. and Zwick, R. (1990) Disentangling the NAEP 1985-86 reading anomaly. Princeton, Educational Testing Service.