Mathematical and Ideological Assumptions in the Modelling of Test Item Responses

Harvey Goldstein
Institute of Education, University of London

INTRODUCTION

The notion that there can be *a theory* of measurement in fields such as education or psychology, is both seductive and elusive. Seductive because it conveys the promise of high scientific status, and elusive because it has so far defied achievement. The present paper examines how psychometric test models based on certain assumptions have come to be used by many practitioners counter-productively and in ways which severely limit the kinds of conclusions which can be drawn.

One does not have to look very far to find claims that a particular statistical model or class of models provides a form of "objective measurement" or constitutes a "theory" of testing; and those claims are not restricted to the purely technical properties of the models, but also carry a substantive message.

The dubious credit for the first serious introduction of the notion of mental traits surely must go to the Galton. It was he who supposed, by analogy with anthropometrical measurements, that there were dimensions of the mind; that there was something called intelligence which was supposed to be normally distributed amongst the population. Much of twentieth century psychometrics has been devoted to elaborating this proposition by developing statistical techniques for ever more complex modelling of supposed mental structures. Item response modelling is merely the most recent example of such activity.

THE ELEMENTS OF ITEM RESPONSE MODELS

A reasonably general model, a so called *two-parameter* binary response model with more than one subject-ability dimension can be written as

$$g(\pi_{ij}) = \sum_{k=1}^{p} \alpha_{kj}\theta_{ki} + \beta_j + \sum_{m=1}^{s} \gamma_m Z_{mi} \qquad (1)$$

Where θ_{ki} is the value for subject i on the k-*th* ability dimension; β_j, α_{kj} are facility and discrimination parameters respectively for the j-*th* item, and the z_{mi} are observed covariates, for example representing group membership. If

$$p = 1,\ \alpha_{kj} = 1,\ g(x) = \text{logit}(x), \text{ and } \gamma_m = 0$$

then we have the notorious Rasch model. If instead $g(x)=x$ then we have the linear *one-parameter* model. This latter model implicitly underlies traditional item analysis procedures such as those for estimating reliability and calculating discriminations. From this point of view the more recent logit item response models are simply a more sophisticated development of the same approach, a point which seems to be poorly understood. More detail on all this can be found in Goldstein and Wood (1989). Models which introduce extra parameters, for example to deal with polytomous responses, do not raise new issues affecting the principles with which this paper is concerned.

Equation (1) expresses the relationship between the probability of a correct response to item j from subject i. In order to be able to provide estimates of the parameters of (1) based on observed data we have to make some statistical assumptions. The key ones are as follows.

DIMENSIONALITY

Consider the simple two -dimensional case of model (1)

$$g(x) = \alpha_{1j}\theta_{1i} + \alpha_{2j}\theta_{2i} + \beta_j \qquad (2)$$

In order to obtain estimates for the parameters of (2) the usual procedure is to regard the θ_{ki} as realisations of random variables Θ_k. An alternative is to regard the θ_{kii} as fixed parameters and then impose sufficient constraints for estimability; an example of this approach is that given by Goldstein (1980). This latter approach is rather difficult to motivate in many practical situations and the former approach suffers from the traditional problems associated with factor analysis models notably the arbitrariness of any particular solution and problems of determining the number of dimensions.

The factor analytic approach has increased in popularity, but has a difficulty that is seldom elaborated. This is the issue of the "reference population". Clearly, it is possible to have a model which "fits" well in one subpopulation, but not in another; for example where the subpopulations comprise different ethnic groups. This problem is well recognised in the ordinary linear model analysis of observational data, and there is a large theoretical and applied literature dealing with it.

While the best psychometric practice does challenge assumptions such that of unidimensionality, this is not currently *standard* psychometric practice. The implications of this assumption is now explored a little further.

The application of unidimensional models often embodies a strong element of tautology, as can be seen in the following example.

Suppose we have the two-dimensional item response model (IRM):

$$\text{logit}(\pi_{ij}) = \alpha_{1j}\theta_{1i} + \alpha_{2j}\theta_{2i} + \beta_j \quad (3)$$

and suppose we assume the commonly used two-parameter unidimensional model

$$\text{logit}(\pi_{ij}) = \alpha_j\theta_i + \beta'_j \quad (4)$$

The estimates of α_j, θ_i are complex weighted functions of the responses and hence of the coefficients in (3). By varying the latter, for example by choosing items with particular values of these coefficients, necessarily we will change the values of α_j, θ_i.

It follows that where a two-dimensional (or more generally multidimensional) structure exists, the choice of items to be included in a test will determine the parameters of a unidimensional model fitted to that structure. Because these parameters are complex functions of the parameters of the underlying multidimensional model, in general they will have no separate interpretation of their own.

If we make the reasonable assumption that life is most commonly multidimensional, then where attempts are made to obtain a unidimensional structure by removing "misfits" etc., any resulting unidimensional model estimates will inevitably reflect the original choice of items. In other words, the test constructor's choice of items to represent what is to be measured is crucial.

Other procedures, such as almost all forms of test equating, which assume unidimensionality, are suspect for these reasons. If there are really several dimensions, and if populations differ along these dimensions, then serious distortions will arise and the techniques will produce results with invalid interpretations.

A similar issue arises with techniques for dealing with item "bias", often referred to as "differential item functioning" (DIF) and sometimes travelling under other labels such as "appropriateness measurement". Briefly, these lead to a test construction procedure whereby items which exhibit extreme or idiosyncratic patterns in terms of group differences typically are marked for exclusion, or at least modification. Thus, if most items on a test discriminated well between men and women, those that didn't would be viewed with suspicion. Shepard et al (1981) sum up this view: "an item is biased if two individuals *with equal ability* but from different groups do not have the same probability of success on the item" (my emphasis).

There is, of course, no way to determine whether "ability" is equal other than from the performance on the test. DIF models assume a particular dependency of item responses on traits and grouping factors. Thus tests for DIF are essentially tests of such a model and alternative hypotheses could include either modifications to the dimensionality structure or to the grouping structure, possibly including further covariates. Thus, DIF studies are perhaps best viewed as exploratory techniques for model fit. Their problem is that they are non-specific since they are concerned with any departure rather than a specific one.

OTHER MODELS

Some authors have begun to question the dominance of current item response models. For example, Masters and Mislevy (1991) advocate latent class models for the allocation of students to "stages" which need not be uniquely ordered. This seems to be an approach worth exploring, especially in diagnostic assessment. Yet like factor analysis it has an attendant set of problems such as determining the number of classes and interpreting the results. It is not clear in their discussion, however, why any particular form of statistical model should be more appropriate for one measurement rather than another. By the same token there is nothing inherently more "theoretical" about these models than the ones

I have already discussed, although the attempt by these authors to start from a substantive problem and then search for an appropriate model is in welcome contrast from a common tendency to apply simple Item Response models to everything in sight.

CONTEXT

A large body of research, especially in mathematics and science, has shown how the contextual embedding of a test question can markedly alter the response success rate. The British Assessment of Performance Unit research has shown how the layout of maths questions changed the response when the actual mathematics content remained the same (Foxman et al, 1990). Murphy (1989) has shown how the practical context of performance on authentic test questions tends to favour girls rather than boys when real life tasks are emphasised as opposed to algorithmic problems. Other research (Wolf et al, 1990) has demonstrated how elusive the notion of "skills" can be when one tries to measure them in practical contexts - a finding which cast considerable doubt upon the current fashion for "criterion referenced" assessment.

In the US the so called National Assessment of Educational Progress (NAEP) reading anomaly (Beaton and Zwick, 1990) demonstrated how sensitive test questions could be to the company they keep. The same small set of questions gave different results depending on how they were embedded in the test. This has led to a search for ways of measuring such effects and that seems to be a fruitful approach. This is, of course, not the same as trying to find context - independent test questions. Such an activity may have its uses, but is not likely to be helpful in most practical situations. Beside the enormous problems of trying to understand the effects of context, the ever more elaborate development of IRM's seems rather irrelevant.

SOCIAL, IDEOLOGICAL AND OTHER CONTEXTS

I have already mentioned the general problem of context influencing performance and now I want to elaborate on some wider implications.

The case of the Golden Rule Insurance company vs. Educational Testing Service (ETS) has been discussed at some length in recent years (Goldstein, 1989, Anrig, 1988). Briefly, the Insurance company managed to persuade ETS to adopt a policy of item selection for its entry test which minimised Black-White differences. It worked by ETS choosing a pool of items, all of which satisfied standard criteria for test inclusion. From this pool the final selection was made by choosing those items which produced the smallest (on average) differences between Blacks and Whites. After some years of this ETS decided that the whole thing had been a mistake and that they wanted to end the agreement.

It was clear from the reaction of many psychometricians to the Golden Rule issue (see, for example Linn and Drasgow, 1987) that they did not want any non-technical considerations to influence the way that tests are constructed. A natural enough reaction from a professional group perhaps, but not necessarily one that society at large would wish to accept. Yet why should we not impose extra-technical constraints on our test constructors?

The idea that tests should seek to minimize or otherwise manipulate group differences is a perfectly legitimate *ideological* aim. It can therefore be attacked ideologically, but it can only be attacked technically if the technical grounds have a theoretical basis. This makes the search for good substantively grounded theory important. Without substantive theoretical support the notion that there is a technical view which can decide this issue (and related ones) is itself an ideological assumption which can and should be challenged. Of course, the search for a theoretical grounding cannot be separated from ideological considerations, but it is better that these are explicit rather than implicit consequences of particular choices of models.

Of course, I am fully aware that by opening up what has previously been a professional process to public participation, there may be all kinds of problems and instabilities and uncertainties which could make life uncomfortable. That however, seems to me to be rather a good thing, although I doubt many in the profession could be persuaded to rally to such a cause. Nevertheless, I perceive the demystification of item response "theory" as a step in the right direction.

REFERENCES

Anrig, G. R. (1988). ETS replies to Golden Rule on "Golden Rule". *Educational Measurement: Issues and Practice; 7*, 20-21

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

Bock, R. D., Gibbons, R. and Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Foxman, D., Ruddock, G. and McCallum, I. (1990). *APU mathematics monitoring 1984-88 (Phase 2).* London; Schools Examination and Assessment Council.

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

Goldstein, H. (1989). *Equity in Testing after Golden Rule.* Institute of Education. (UnPub)

Goldstein, H. and Wood, R. (1989). Five decades of item response modelling. *British Journal Mathematical Statistical Psychology, 42*, 139-167.

Gould, S. J. (1981). *The Mismeasure of Man*, New York: W. W. Norton.

Linn, R. J., and Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement: Issues and practice, 6*, 13-17.

Murphy, P. (1989). Assessment and gender. *National Union of Teachers Education Review, 3*, 37-41.

Masters, G. N. and Mislevy, R. J. (1991). New views of student learning: implications for educational assessment. In Frederiksen, N., Mislevy, R. J. and Bejar, 1. 1. (eds.); *Test theory for a new generation of tests.* New Jersey, Lawrence Earlbaum Associates.

Shepard, L., Camilli, G. and Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. *Journal of Educational Statistics 6*, 317-375.

Wolf, A., Kelson, M. and Silver, R. (1990). *Learning in Context: patterns of skills transfer and training implications*, London: The Training Agency.