70

British Journal of Mathematical and Statistical Psychology (1980), 33, 234–246 ***Printed in Great Britain***

Dimensionality, bias, independence and measurement scale problems in latent trait test score models

Harvey Goldstein

It is argued that latent trait models used in the analysis of test scores, in particular the Rasch model, suffer from serious defects. They assume implicitly a measurement scale model for which little substantive justification seems to be available and they make the assumption of 'local independence' which enjoys little empirical support. Perhaps their most serious drawback, which is illustrated with the 'fixed effects' Rasch model, is that the individual ability estimates obtained are always biased. This bias is not constant but depends upon an individual's true ability, the number of items in the test and their difficulty values. Moreover, there is no way to obtain an unbiased estimate in the typical case when only one administration of a test is given to individuals. This result suggests that the use of such models needs to be critically examined, especially in connection with so-called 'item banks'. The paper also discusses the dimensionality of the latent space, and shows how multidimensional models may be specified.

1. Introduction

This paper investigates some mathematical and statistical properties of certain latent trait test score models, which seem to have been largely ignored or to have suffered from imprecise specification. The applicability of these models in substantive areas such as education has been discussed elsewhere (see Goldstein & Blinkhorn, 1977; Goldstein, 1979) and will not be discussed further here, although naturally the results of the present paper have clear implications for such applications. In the first section the measurement model underlying the commonly used logistic model is studied. The second section examines the usual assumption of local independence and its relation to the dimensionality of ability, and the third section presents a method for fitting models with two or more ability dimensions. The final section demonstrates that the usual estimates for the values of the ability parameters are inherently biased.

Throughout the paper, the simple logistic or logit model known as the Rasch model (Rasch, 1960) will form the basis for discussion. This is partly because it is mathematically relatively easy to analyse, but largely because it has become one of the most popular latent trait models used in the design and analysis of mental tests and has accumulated a sizable literature. Nevertheless, the results of the following sections are also generally applicable to the very similar normal ogive (or probit) model, and can readily be adapted to more complex models such as those which specify more than one parameter for each item in order to allow, say, for differing discriminating powers or for guessing.

2. Measurement scale models

The Rasch model can be written

$$\text{logit}(P_{ij}) = \log[P_{ij}/(1-P_{ij})] = \alpha - \beta_j + \delta_i \quad (1)$$

with an arbitrary linear side condition on the β_j and one on the δ_i and where $j = 1, \ldots, m$ refers to individuals, and $i = 1, \ldots, k$ refers to items. Thus P_{ij} is the probability of the jth individual responding correctly to the ith item.

In a typical realization, a test consisting of k binary items is administered just once to each individual. This paper is primarily concerned with the **'fixed effects' version** of (1) where each β_j, δ_i is regarded as a parameter to be estimated. **An alternative** model which has received some attention is the 'mixed effects' model where β_j is assumed to be random and to have a normal distribution in the population (see, for example, Sanathanan & Blumenthal, 1978 and also Bartholomew, 1980). This model avoids certain difficulties associated with the fixed effects model, but it does make, typically, the assumption of normality, and it would also seem to be unsuitable for many applications. Some relevant properties of the mixed effects model will be referred to later.

The basic data matrix resulting from the application of a test to a sample of individuals is a three-way contingency table of individuals by items by response, with the last factor having just two levels: success or failure. Model (1) treats the response factor as the dependent variable relating the probability of a correct response (success) to an additive function of individual ability (β_j) and item difficulty (δ_i). For a given item i and individual with ability β we can write the probability of a correct response as

$$P_\beta = A\,e^\beta(1 + A\,e^\beta)^{-1}, \quad \text{where } A = \exp(\alpha + \delta_i). \tag{2}$$

The form of this relationship (setting the factor $A = 1$ for convenience) is shown in Fig. 1 by the continuous line. This is symmetrical about $P = \frac{1}{2}$, $\beta = 0$, and as $P \rightarrow 0$ or 1, so $\beta \rightarrow -\infty$ or $+\infty$. Thus, for large (or small) values of P, to achieve a small change in the probability of success requires a large change in ability. For example, the difference in ability between an individual with 99 per cent probability of success and one with a 95 per cent probability of success is about 1·6 units of ability, which is the

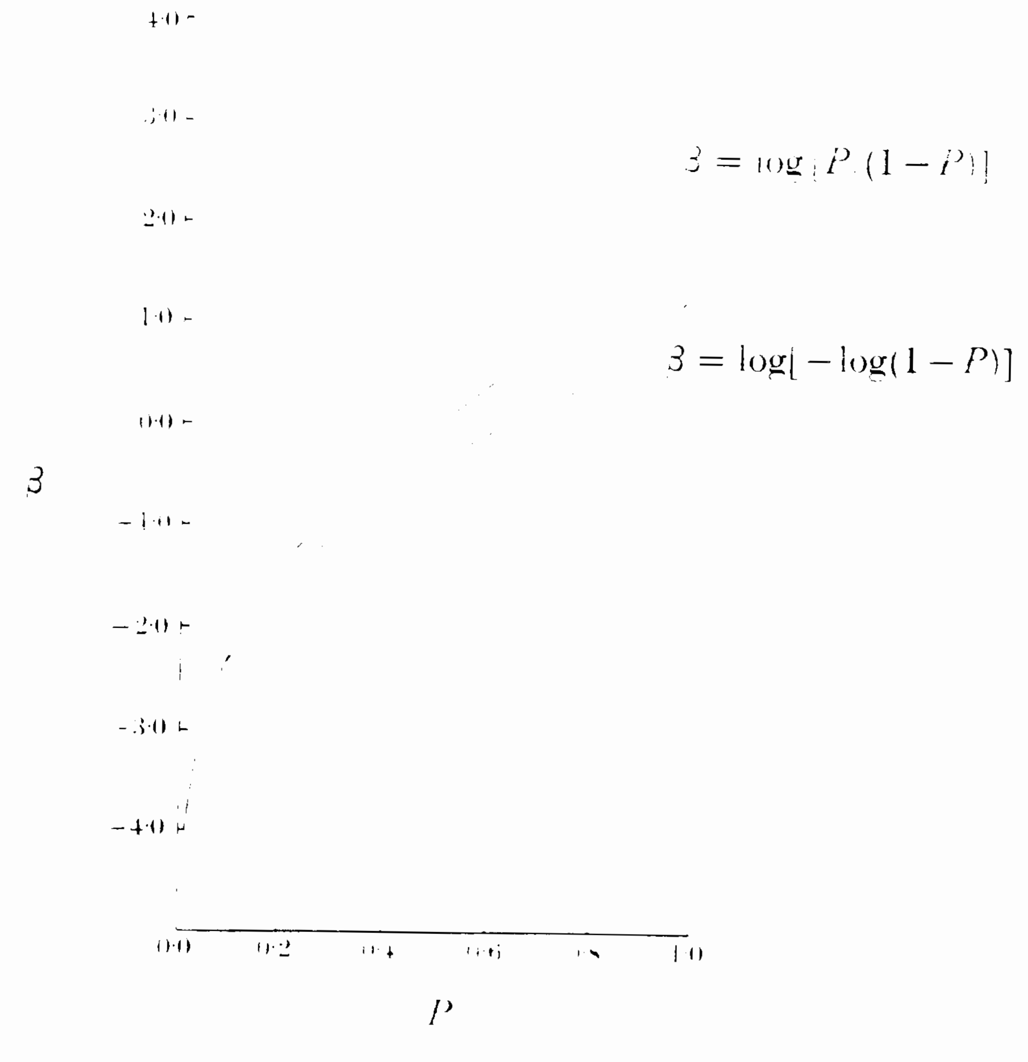

Figure 1. Illustration of the general relationship between latent ability β and the probability of a correct response P to a test item for the logit and log-log response models.

same as the difference in ability between an individual with a 30 per cent probability of success and one with a 70 per cent probability of success. Thus the model discriminates far better at the extremes of the ability scale than in the middle.

The measurement scale properties of the model are also apparent where ability estimates are found from a set of items comprising a test. A typical example is given by Cohen (1979) who provides a table for a 50-item physics test where the difference in estimated ability between those scoring 48 out of 50 correct responses and those scoring 49 out of 50, about 1·6 per cent of the sample, is 0·73 units. This is approximately equal to the difference between those scoring 21 out of 50 and those scoring 29 out of 50, this range including about 36 per cent of the sample. Such a scaling procedure seems to be quite arbitrary and it is difficult to see any substantive justification for it.

In the fixed effects model, individuals who score all successes or all failures cannot be assigned finite ability estimates. In the mixed effects model the assumption of an underlying normal distribution for the ability parameter β does enable finite estimates of β to be obtained for these individuals by evaluating $E(\beta|r, \mu, \sigma)$ where r, the raw score, is the number of correct responses, and μ, σ are the (estimated) mean and standard deviation of the distribution of β. From a measurement point of view, however, the assumption that ability has a normal distribution in a population can be queried, and alternative distributional forms will in general provide different estimates.

There is surprisingly little discussion in the literature about the measurement scale properties of model (1). For example, Lord & Novick (1968) in their standard text on mental measurement theory provide no adequate discussion of these properties. Ferguson (1942) seems to have been one of the first to introduce the closely related normal ogive latent trait test model, but he fails to provide an adequate psychological justification for the underlying measurement scale. There also seems to have been an early failure to distinguish clearly between the assumption of a normal ogive shape for the item characteristic curve, the so-called 'phi-gamma hypothesis' (Guilford, 1936), and the assumption of an underlying normal distribution of the ability parameter in the population. The subsequent development of the normal ogive and logistic models by Lawley (1943) and others, stemming from Ferguson's work, seems to have spent little time considering the measurement scale issue. It is true of course that models where β tends to $\pm\infty$ as $P \to 0$ or 1 are convenient, since the efficient estimation of the parameters of the equation

$$g(P_{ij}) = \alpha + \beta_j + \delta_i,$$

where $g(0)$ and $g(1)$ are finite, would in general involve complicated constrained minimization techniques, as well as posing problems of interpretation.

It might be argued that if the measurement scale properties of model (1) were undesirable, nevertheless the resulting ability estimates could always be transformed non-linearly to produce a substantively more acceptable scale. While this is possible, it seems rarely to be done, and if one is subsequently going to carry out a non-linear transformation of ability it is difficult to see why the original model should have been used in the first place.

To illustrate the effect of adopting a model with somewhat different scale properties, a set of data has been analysed fitting first the Rasch model and secondly the following log–log model:

$$\log[-\log(1 - P_{ij})] = \alpha + \beta_j + \delta_i,$$
$$P_{ij} = 1 - \exp[-\exp(\alpha + \beta_j + \delta_i)], \qquad (3)$$

where the parameters have the same meanings as before.

Corresponding to equation (2) we have

$$P_\beta = 1 - \exp(-A \exp \beta). \tag{4}$$

As with the Rasch model, the item characteristic curves for this model are non-intersecting and $\beta \rightarrow \mp \infty$ as $P \rightarrow 0$ or 1 for any item. The broken line in Fig. 1 shows the relation (4) with $A = 1$. For small values of P this curve is indistinguishable from that for the Rasch model, but as $P \rightarrow 1$ the log–log function approaches ∞ much more slowly than the logit function. Thus, as a measurement scale, the log–log transformation of P does not lead to the same degree of 'stretching' at the high ability end of the scale. It could be argued that it is therefore a more suitable model for those applications, such as the physics test quoted above, where we wish to avoid an exaggeration of small differences in success rate for large values of P. It might also be argued that the 'stretching' of the scale for low values of P is justified on the grounds that there ought to be a large difference in ability between an individual whose average success rate was, say, 2 per cent and one whose rate was 4 per cent. Hence, this would argue that the measurement scale for β should not be symmetrical with respect to correct and incorrect responses. These remarks are not intended to constitute a general argument in favour of the log–log as opposed to the logit model, but merely to show that from a measurement scale viewpoint the logit model may have deficiencies which merit serious attention.

To illustrate some further differences, a comparison of these two models will be made using some mathematics test data. Nine items have been selected from an 'O' level General Certificate of Education examination paper containing 60 items in all. The chosen items are all concerned with two-dimensional Euclidean geometry and the responses of 46 16-year-old children are analysed. (Four children were excluded from an original sample of 50 because they had answered all items correctly and thus could not be assigned finite values in either model.) All the parameters in models (1) and (3) are estimated simultaneously by maximum likelihood, assuming a binomial error distribution and local independence. Following a common practice, the parameter α is absorbed into β_j and the side condition on the parameters δ_i is that they sum to zero. Unlike the logit model where the raw score (number of correct responses) is a minimal sufficient statistic for an individual's ability parameter, no minimal sufficient statistic exists for the log–log model. For the latter model, therefore, a conditional maximum-likelihood procedure (Anderson, 1970), which allows the item parameter values to be estimated without simultaneously estimating the ability parameters, is not available, and it is known that the unconditional maximum-likelihood estimates of the parameters as used here, are inconsistent and biased. Wright & Douglas (1977) suggest that the unconditional maximum-likelihood estimates of the item parameters in the Rasch model are too large by a factor of approximately $k/(k-1)$, with presumably a similar result holding for the ability parameters. Since the purpose of the present analysis is to compare the general behaviour of two measurement scales, such biases will be unimportant.

Table 1 gives the individual ability parameters corresponding to each raw score $r = 1, \ldots, 8$. For the logit model, since r is sufficient, every individual with a score of r has the same estimated value of β. For the log–log model this is not the case and the median value of β is shown, together with each individual value, for each raw score group.

Two points are worth noting about this table. First, for each raw score, the log–log model has quite a large range of variation of parameter values reflecting the number of different response patterns giving rise to the same score. For example, the individual

Table 1. Individual ability estimates for the logit and log–log models for the set of nine geometry items

		Parameter estimates		
			Log–log model	
Raw score	Number in sample	Logit model	Median	(Individuals)
2	4	−1·86	−1·90	(−1·94, −1·93, −1·86, −1·86)
3	3	−0·95	−1·39	(−1·87, −1·39, −1·33)
4	6	−0·25	−0·78	(−0·90, −0·83, −0·78, −0·78, −0·75, −0·75)
5	8	0·42	−0·46	(−0·62, −0·53, −0·50, −0·48, −0·43, −0·39, −0·31, −0·26)
6	4	1·11	0·08	(−0·26, 0·03, 0·14, 0·19)
7	8	1·92	0·51	(0·28, 0·49, 0·43, 0·43, 0·59, 0·70, 0·70, 0·70)
8	13	3·08	1·44	(0·23, 0·93, 1·10, 1·10, 1·44, 1·44, 1·44, 1·44, 1·44, 1·44, 1·44, 1·44, 1·44)
Goodness-of-fit χ^2 (360 d.f.)		319·9	315·0	

with ability estimate 0·23 and a raw score of 8 has been relatively 'penalized' by giving the incorrect answer to the easiest of the items. Thus the log–log model allows a weighting for each item, related to its difficulty value, to enter the estimation of the parameters. Whether this is desirable will depend upon the particular application but, for example, where there are a few difficult items, one might well wish to attach a relatively greater weight to a correct response to these than to the other items. The existence of a minimal sufficient statistic for the logit model, while statistically attractive, does not imply necessarily that the model therefore has useful or desirable measurement scale properties. Secondly, the overall goodness of fit statistics for both models are similar and below expectation, despite the differences in parameter estimates. This underlines the point that the value of a formal test statistic is never in itself sufficient justification for accepting a particular model, apart from the fact that, with zero-one data and a relatively small sample, the distribution of the goodness-of-fit statistic cannot be expected to approximate closely to a χ^2 distribution. There are other tests of fit which are more powerful against specific alternatives (see, for example, Anderson, 1973), but in general the same point still applies.

3. Local independence

Both the models so far considered assume that the zero-one responses in the basic data matrix are statistically independent. In most contingency table applications this assumption is satisfied because each response is contributed by just one independently selected individual member of a random sample. In the case of mental tests, however, only the sets of responses between individuals can be made independent in this way, and for responses within individuals there exists the possibility of mutual dependencies. For example, for any individual the probability of a correct response to item j may depend upon whether the response to item $j-1$ was correct or not. The formal requirements for independence of item responses within an individual may be written as

$$f(y_1, \dots, y_k \mid \beta) = \prod_{i=1}^{k} g(y_i \mid \beta), \tag{5}$$

where $f(y_1, \ldots, y_k|\beta)$ is the joint distribution of responses $y_1, \ldots, y_k$ for a given value of ability β, and $g(y_i|\beta)$ is the marginal distribution for the ith item given β. In the general multidimensional case β represents a vector of parameters.

The $y_1, \ldots, y_k$ (zero-one) responses themselves form a 2^k contingency table with each item corresponding to a way of classification and (5) says that this table can be described in terms of the margins only. Models (1) and (3) specify that the k parameters fitted to these margins are the same for each individual. If this is not the case then 'interaction' terms are required in these models. In the extreme case each individual will require a different set of parameter values, so that a separate parameter will be needed for each of the $k \times m$ cells of the table, namely the same number as the number of available (zero-one) observations.

If it were possible to apply the test many times to the same individuals, then the cells of the total $m \times 2^k$ table could be filled in, making it possible to test for dependencies among items and to fit models incorporating various dependency patterns. Within a single application, however, there is no general method available for studying item dependencies. It should also be noted that the joint distribution of item responses, $f(y_1, \ldots, y_k|\beta)$, is defined with respect to a test in which the conditions of administration, for example the order of presentation of the items, is fixed. If, say, the item ordering was changed, then any dependencies between items and hence the joint distribution might also change.

There may be some kinds of dependencies which can be tested indirectly. For example, if we deliberately present a set of items in different orders to random samples, or present samples with different mixtures of items, any significant differences in estimated parameter values might be interpreted as resulting from a lack of local independence, although other interpretations would also be possible. There seems to have been little systematic attempt to carry out suitable experiments, or to study the consequences for estimation and inference procedures when assumption (5) is violated. Without the results of such studies it is difficult to be sure how serious might be any failure of (5). Nevertheless, any dependencies among item responses will invalidate the present methods of analysis which assume (5) to be true, and would lead, for example, to incorrect goodness-of-fit tests. In fact, the assumption of local independence is such a strong assumption that it would be surprising if it were true other than in a few specially contrived circumstances.

One further point concerning local independence is worth mentioning since there seems to be confusion over its relationship to the dimensionality of the between-individuals latent space. In (5) the vector β of length q, say, specifies this latent space, and each individual will be represented by a point in this q-dimensional space. Thus, if we have two individuals with the same values of β, their joint distribution of item responses, $f(y_1, \ldots, y_k|\beta)$, will be identical. It should be noted, however, that the dimensionality of β can be specified without invoking the assumption of local independence. We can, for example, have a one-dimensional model such as the logistic, either with or without local independence. Lord & Novick (1968, p. 361) claim that 'The assumption of local independence is thus equivalent to the assumption that θ (i.e. β) spans the complete latent space'. An examination of their argument in support of this claim, however, shows that they have failed to distinguish properly between (a) the conditional distribution of the item responses for a given β and (b) the conditional distribution of the item responses for a given β *and* given responses to other items. This arises partly from their definition (p. 359) of the latent space in terms of a single item response distribution rather than in terms of the joint distribution of item responses.

This point is of some practical importance since other authors (for example, Gustafsson,

1980) have used the Lord & Novick claim to avoid testing the assumption of local independence once they had settled the dimensionality of the latent space.

4. Dimensionality of the latent space

We now turn to the detailed consideration of the dimensionality of the between-individual latent space for fixed effects models. For convenience the Rasch model will be studied, although similar conclusions apply to other models such as (3).

Model (1) supposes a one-dimensional latent space since all differences between individuals are summarized by a single variable. Moreover, it is not possible to specify any further dimensions which are independent of the item set, since the parameters of these dimensions would then have to be specified by terms fitted to the margin referring to individuals, in the items × individuals table, and we have already fitted the maximum possible number of parameters (m) to this margin in order to specify the first dimension. Hence, if we wish to incorporate further parameters in to (1) they must be components of the individuals × items interaction. These remarks do not apply to the mixed effects model where more than one random dimension between individuals could be defined, but there seems to have been little development of such models (but see Bartholomew, 1980).

In spite of the foregoing remarks it is still possible to study aspects of dimensionality. Suppose a set of k items is divided into two subsets A and B, consisting of k_1 and k_2 items respectively and suppose that this is done on the basis that these two sets are related to different underlying psychological or educational dimensions. Then we can write the following model:

$$\operatorname{logit}(P_{ij}) = \beta_j + \delta_i + \gamma_{Aj} + \gamma_{Bj}, \tag{6}$$

where

$$\gamma_{Aj} = V_j \quad \text{if } i \in A, \quad 0 \text{ otherwise},$$

$$\gamma_{Bj} = W_j \quad \text{if } i \in B, \quad 0 \text{ otherwise}$$

and α has been absorbed into β_j.

Thus, the response of the jth individual is determined by two quantities, one from set A and one from set B so that we have a two-dimensional model. If $\gamma_{Aj} = \gamma_{Bj}$ then, of course, we are back with the one-dimensional model. The extension to three or more dimensions is obvious. If we can classify items into different sets reflecting substantively different underlying dimensions, then we can estimate the relevant parameters and also test for the existence of a lower number of dimensions.

The likelihood of (6) can be written as

$$\begin{aligned} \Lambda &= \prod_j \prod_i \{[\exp(\beta_j + \delta_i + \gamma_{Aj} + \gamma_{Bj})]^{a_{ij}}[1 + \exp(\beta_j + \delta_i + \gamma_{Aj} + \gamma_{Bj})]^{-1}\} \\ &= \prod_j \prod_{i \in A} \{[\exp(\beta_j + \gamma_{Aj} + \delta_i)]^{a_{ij}}[1 + \exp(\beta_j + \gamma_{Aj} + \delta_i)]^{-1}\} \\ &\quad \times \prod_j \prod_{i \in B} \{[\exp(\beta_j + \gamma_{Bj} + \delta_i)]^{a_{ij}}[1 + \exp(\beta_j + \gamma_{Bj} + \delta_i)]^{-1}\} \\ &= \Lambda_A \Lambda_B, \end{aligned}$$

where $a_{ij} = 1$ if the response to item i for individual j is a success and 0 otherwise, and Λ_A and Λ_B are the separate likelihoods for the two sets A, B. Hence the parameters of (6) can be estimated from the separate analyses of the two sets and a test for $\gamma_{Aj} = \gamma_{Bj}$ is obtained by comparing goodness-of-fit X^2 statistics. A similar result holds for the log–log and probit models.

The procedure is illustrated using the same data as before, with the addition of four algebra items. The set A thus consists of geometry items with $k_1 = 9$ and the set B of algebra items with $k_2 = 4$.

Model (1) was first fitted to all 13 items and yielded the goodness-of-fit $X^2 = 541{\cdot}5$ with 540 degrees of freedom. For the set A the goodness-of-fit $X^2 = 319{\cdot}9$ with 360 degrees of freedom and for set B $X^2 = 148{\cdot}7$ with 135 degrees of freedom. None of these values is significant at the 10 per cent level. The test for model (1) against model (6) gives $X^2 = 541{\cdot}5 - (148{\cdot}7 - 319{\cdot}9) = 72{\cdot}9$ with 45 degrees of freedom which is significant at the 1 per cent level, indicating the existence of two dimensions. For the log–log model the corresponding test statistic has a similar value, $X^2 = 73{\cdot}7$. Given such a result we would normally wish to compare the parameter estimates for sets A and B, for example in a scatterplot. In the present case this would not be very informative due to the small number of items involved, especially in set B. The existence of an apparently adequate fit for all 13 items as well as a highly significant test for a second dimension, further underlines the point made in Section 2, that the value of a formal test statistic is never in itself sufficient justification for accepting a particular model.

5. Estimation bias

It was noted in Section 2 that, in the fixed effects logit and log–log models, items or individuals where the responses were either all successes or all failures could not be assigned finite parameter estimates and had to be omitted from the analysis. We shall assume in what follows that the item parameter values are known and are to be used to estimate the individual ability parameters, which is what happens in the two-stage approach of the conditional maximum-likelihood method for the logit model. Nevertheless, where items are excluded from an analysis because the sample individuals either all answer an item correctly or all answer incorrectly, then we can carry out a similar analysis for items to that given below for individuals. We shall work with the fixed effects model (1), and assume that it holds, together with local independence. Similar arguments to those which follow can be applied to model (3) also.

Consider again the two-way table with cell entries consisting of ones or zeros. In (1) P_{ij} is the probability of a success in cell (i, j) and can be thought of as the limiting relative frequency of successes resulting from a notional repeated application of item i to individual j. In practice, of course, we normally can do this only once, but nevertheless the aim of any analysis of a set of item responses is to provide an estimate of these underlying probabilities.

In the notional repeated applications of a set of k items to an individual, some applications will, by chance, result in a complete set of failures or a complete set of successes. The probabilities of these events are respectively

$$P_\beta(0) = \prod_{i=1}^{k} [1 + \exp(\beta + \delta_i)]^{-1}$$

$$P_\beta(k) = \prod_{i=1}^{k} [\exp(\beta + \delta_i)][1 + \exp(\beta + \delta_i)]^{-1}.$$

Clearly, as β increases so $P_\beta(k)$ increases and $P_\beta(0)$ decreases. In practical applications, a set of items is only administered once and, as pointed out earlier, if the responses are all successes or all failures no finite estimate of β is possible. If $\hat{\beta}$ is the estimated value of β it is easy to see that the expected value of $\hat{\beta}$, averaged over notional repeated applications *where instances with total successes or failures are excluded*, will in general not be equal to β.

For the fixed effects model we write $P(r|\delta, \beta)$ for the probability of obtaining raw score r out of k for the set of item parameter values δ and the ability value β $(r = 1, \ldots, k-1)$. Let $\hat{\beta}_r$ be the usual maximum-likelihood estimate of β for an obtained raw score r, found by solving the equation

$$r = \sum_{i=1}^{k} [\exp(\beta - \delta_i)][1 + \exp(\beta - \delta_i)]^{-1}. \tag{7}$$

It can be shown that

$$P(r|\delta, \beta) = \exp(r\beta) A_r \left\{ \prod_{i=1}^{k} [1 + \exp(\beta - \delta_i)] \right\}^{-1}, \tag{8}$$

where

$$A_r = \sum_{\Sigma a_i = r} \prod_{i=1}^{k} \delta_i^{a_i},$$

which is an augmented symmetric function of the δ_i. The required expected value of $\hat{\beta}$ is

$$E(\hat{\beta}|\beta) = \left[\sum_{r=1}^{k-1} \hat{\beta}_r P(r|\delta, \beta) \right] \left[\sum_{r=1}^{k-1} P(r|\delta, \beta) \right]^{-1}, \tag{9}$$

which can thus be calculated for any value of β from a knowledge of the values of the item parameters δ_i. It is worth noting that, for $r = 1$ or $r = k-1$, the solution to (7) corresponds to $\hat{\beta} = \pm\infty$ in (9). The behaviour of the bias $E(\hat{\beta}|\beta) - \beta$ for different values of β will now be studied using the item parameter values δ_i obtained from the nine geometry items used earlier. A similar analysis can be carried out for the log–log and probit models with the same general conclusions, although the absence of a minimal sufficient statistic complicates the computations.

Table 2 gives the estimates of the δ_i, together with the value of $\hat{\beta}_r$ for $r = 1, \ldots, 8$. The δ_i values are those obtained from the unconditional maximum-likelihood analysis described earlier. Since this example is only illustrative, it is unimportant whether a conditional or unconditional procedure is used, although in a real application it would

Table 2. Item difficulty estimates and resulting individual ability estimates for the set of nine geometry items, for the logit model

Item (i)	Parameter estimate (δ_i)	Raw score (r)	Parameter estimate ($\hat{\beta}_r$)
1	3·81	1	−3·24
2	0·12	2	−1·86
3	1·62	3	−0·95
4	−0·04	4	−0·25
5	−0·89	5	0·42
6	−0·47	6	1·11
7	0·60	7	1·92
8	−1·57	8	3·08
9	−3·20		

be better to use the conditional estimates. Table 3 gives values of $P(r|\delta,\beta)$ for $r=0$ and k (here $k=9$), $E(\hat{\beta}|\beta)$ and the bias $E(\hat{\beta}|\beta)-\beta$ for a range of values of β. As k becomes large, so $P_\beta(0)$ and $P_\beta(k)$ will tend to become smaller for a given value of β and the bias will also tend to decrease. Table 4 gives the value of the bias for tests of 18, 27 and 36 items which are formed from the original nine items with the parameter values repeated twice, three and four times respectively.

Tables 3 and 4 show clearly that for small values of k, the bias can be large even for moderate values of β. The question naturally arises as to whether it is possible to apply a correction to the estimated values $\hat{\beta}_r$ to give unbiased estimates. Suppose, for a test with k items that for $r=1,\ldots,k-1$ we can determine unbiased estimates denoted by β_r'. For a true ability value β the expected value of β_r' is given by

$$\beta' = \left\{ \sum_{r=1}^{k-1} \beta_r' \, P(r|\delta,\beta) \right\} \left\{ \sum_{r=1}^{k-1} P(r|\delta,\beta) \right\}^{-1}. \tag{10}$$

Table 3. Expected values of maximum-likelihood ability estimates, biases and probabilities of total success or total failure for the set of nine geometry items, for a range of true ability values

β	$P_\beta(0)$	$P_\beta(k)$	$E(\hat{\beta}\|\beta)$	Bias $E(\hat{\beta}\|\beta)-\beta$
−5·0	0·72	0·00	−3·16	1·84
−4·0	0·46	0·00	−3·03	0·97
−3·0	0·19	0·00	−2·71	0·29
−2·0	0·04	0·00	−2·07	−0·07
−1·0	0·00	0·00	−1·11	−0·11
0·0	0·00	0·00	−0·02	−0·02
1·0	0·00	0·01	1·06	0·06
2·0	0·00	0·06	2·01	0·01
3·0	0·00	0·26	2·61	−0·39
4·0	0·00	0·56	2·89	−1·11
5·0	0·00	0·79	3·01	−1·99

Table 4. Biases for tests based on 18, 27 and 36 geometry items for a range of true ability values

	Bias $(E(\hat{\beta}\|\beta)-\beta)$		
β	$k=18$	$k=27$	$k=36$
−5·0	0·91	0·51	0·29
−4·0	0·27	0·04	−0·05
−3·0	−0·06	−0·09	−0·08
−2·0	−0·11	−0·07	−0·05
−1·0	−0·07	−0·04	−0·03
0·0	−0·05	0·00	−0·01
1·0	0·02	0·03	0·02
2·0	0·08	0·06	0·05
3·0	0·01	0·08	0·09
4·0	−0·41	−0·13	0·00
5·0	−1·14	−0·74	−0·49

If this is to be unbiased we require $\beta' = \beta$. Equation (10) becomes therefore

$$\beta = \left\{ \sum_{r=1}^{k-1} \exp(r\beta) A_r \beta_r' \right\} \left\{ [1 - P_\beta(0) - P_\beta(k)] \prod_{i=1}^{k} [1 + \exp(\beta + \delta_i)] \right\}^{-1}$$

which gives

$$\sum_{r=1}^{k-1} [\exp(\beta)]^r A_r \beta_r' = \beta \left\{ \prod_{i=1}^{k} [1 - \exp(\beta) \exp(\delta_i)] - \prod_{i=1}^{k} [1 + \exp(\beta) \exp(\delta_i)] - 1 \right\}.$$

Both sides of the equation are polynomials of degree $k-1$ in $\exp(\beta)$. The coefficient of each power of $\exp(\beta)$ on the right-hand side, however, includes β, whereas those on the left-hand side do not. Hence this equation cannot be true for all β and the β_r' can be chosen in general to satisfy (10) for at most $k-1$ values of β.

Another possibility would be to assign values, say b_0 and b_k respectively, for the cases $r=0$, and $r=k$ in order to obtain unbiased estimates. We then have, using the same notation as before,

$$\beta' = \sum_{r=1}^{k-1} \beta_r' P(r \mid \delta, \beta) + b_0 P_\beta(0) + b_k P_\beta(k). \tag{11}$$

For unbiasedness we require $\beta' = \beta$ and this leads to

$$\beta \prod_{i=1}^{k} [1 + \exp(\beta + \delta_i)] = \sum_{r=1}^{k-1} [\exp(\beta)]^r A_r \beta_r' + b_0 + b_k \prod_{i=1}^{k} [\exp(\beta) \exp(\delta_i)].$$

Both sides of this equation are polynomials of degree k in $\exp(\beta)$. Similarly to before, however, the coefficients of each power of $\exp(\beta)$ on the left-hand side include β whereas those on the right-hand side do not, so that this equation can be satisfied, in general, for b_0, b_k and the β_r' for at most k values of β.

Thus there is, in general, no set of values corresponding to $r = 0, 1, \ldots, k$ which will give unbiased estimates of β. It was indicated earlier that item parameter estimates will also be biased where items are excluded on the basis of a set of totally correct or totally incorrect responses from the individuals in a sample. In theory of course, the sample size can be increased so that the probability of exclusion becomes small enough to ensure negligible biases. In practice, however, this may often not happen and there will then be biases which, in general, will depend on the size and composition of the sample.

6. Discussion

The literature on latent trait test score models seems largely to have ignored some fundamental questions concerning the statistical and substantive validity of these models. It has been argued in this paper that the measurement scales implied by such models are questionable in terms of any underlying psychological or educational theories. This is underlined by a comparison of a two-parameter log–log model with the Rasch model. The two models have many of the same features, including the same number and type of parameters, and a constant ordering of item difficulty for all individuals. In addition, both models produce similar X^2 test statistic values when applied to an illustrative data set. Nevertheless, the models give quite different parameter estimates, reflecting the difference in the underlying measurement scale assumptions. The Rasch model has the statistically useful property of possessing a singly sufficient statistic, namely the raw score, but it is by no means clear that it is an advantage for individuals with the same score to be given the same estimate of ability, and it could be argued that other features of the total response pattern should be acknowledged, as in the log–log model. There seems to be a case for giving much more attention to such measurement scale issues.

The second issue concerns another rather neglected topic, local independence. While the assumption of local independence is necessary for the application of standard statistical procedures, it is far from being self-evidently satisfied, and the lack of studies designed to test this assumption is a serious omission. Nevertheless, there are real difficulties involved in testing this assumption. Given the practical constraint that a test of k items typically can only be administered once to any individual, the assumption of local independence can only be tested, if at all, by studying the responses to a set of items administered, for example, in different orders to comparable samples. While this procedure may show the existence of dependencies its results will be open to alternative explanations for changes in difficulty values. It is also shown that local independence is a quite separate assumption from that of the dimensionality of the between-individuals latent space.

The third topic deals with methods for studying the dimensionality of the latent space using straightforward procedures, although, of course, the various problems discussed in the remainder of the paper also apply to models with more than one dimension. It illustrates in particular, that a satisfactory overall 'goodness-of-fit' test for a one-dimensional model is not sufficient justification for concluding that no further dimensions or interactions are present.

The final, and arguably the most important, section demonstrates that for the fixed effects model, individual ability estimates are biased. Also, item difficulty estimates will be biased unless the sample size can be made sufficiently large to avoid very easy or very difficult items being excluded from the analysis. These results follow from the practical constraint that a test of k items typically can only be administered once to any individual, and because the underlying measurement scale extends to $\pm\infty$. The bias in the individual ability estimates tends to decrease as the number of items increases, and there is no way of obtaining unbiased estimates. It could be argued that, for a fixed test, since the bias is constant for a given true ability, the biased estimate may be viewed simply as a particular transformation of the ability scale. It is difficult to see in this case, however, why a given latent trait model is used in the first place since the biased estimates do not conform to the model. Also, a major justification for the use of the Rasch model in particular (Wright, 1977) is that individuals can be assigned scores on a common scale using a wide variety of tests. As has been shown, such a procedure does not lead to the expected results. It would appear, in particular, that the use of item banks based on the Rasch model needs to be examined critically.

While this paper has been principally concerned with the fixed effects model it should be pointed out that some of the problems may be overcome with the mixed effects model. Most importantly, unbiased ability estimates can be obtained, but they are only available for an individual or a group of individuals where these can be regarded as randomly selected from a population with a known distribution of ability.

Acknowledgements

I am grateful to Jan-Eric Gustafsson for the use of a routine to calculate values of the symmetric functions, and to Bob Wood for permission to use the illustrative data set. I am particularly grateful to Steve Blinkhorn, Russell Ecob, Anne Hawkins, Michael Healy, Desmond Nuttall, Ian Plewis and Bob Wood for their very helpful comments on an early draft of this paper.

References

Anderson, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, **32**, 281–301.

Anderson, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, **38**, 123–140.

Bartholomew, D. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, **42** (in press).

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology*, **32**, 113–120.

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika*, **7**, 19–29.

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

Goldstein, H. & Blinkhorn, S. (1977). Doubts about item banking. *Bulletin of The British Psychological Society*, **30**, 309–311.

Guilford, J. P. (1936). *Psychometric Methods*. New York and London: McGraw-Hill.

Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, **33**, 205–233. (Paper presented to Annual meeting of American Educational Research Association, San Francisco, 8–12 April 1979.)

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, **61A**, 273–287.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: The Danish Institute for Educational Research.

Sanathanan, L. & Blumenthal, S. (1978). The logistic model and estimation of latent structures. *Journal of the American Statistical Association*, **73**, 794–799.

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

Wright, B. D. & Douglas, G. A. (1977). Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement*, **37**, 47–60.

Received 30 September [illegible]; revised version received 30 January 1980

Requests for reprints should be addressed to Harvey Goldstein, Department of Statistics and Computing, University of London Institute of Education, 20 Bedford Way, London WC1H 0AL.