DOCUMENT RESUME

ED 228 267 TM 830 019

AUTHOR Choppin, Bruce
TITLE A Fully Conditional Estimation Procedure for Rasch Model Parameters.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO CSE-R-196
PUB DATE Nov 82
NOTE 34p.; Some tables may be marginally legible due to small print.
PUB TYPE Reports - Research/Technical (143) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Difficulty Level; *Estimation (Mathematics); Goodness of Fit; *Item Analysis; *Latent Trait Theory; *Mathematical Models; Matrices; *Maximum Likelihood Statistics; Statistical Analysis; Test Items
IDENTIFIERS *Ability Parameters; Item Calibration; *Item Parameters

ABSTRACT

A strategy for overcoming problems with the Rasch model's inability to handle missing data involves a pairwise algorithm which manipulates the data matrix to separate out the information needed for the estimation of item difficulty parameters in a test. The method of estimation compares two or three items at a time, separating out the ability parameters of the set of persons tested by means of conditional probability, to avoid biasing the difficulty parameters. To describe the difficulty of a set of items, a matrix is constructed in which each element is the number of people who responded correctly to one item and incorrectly to another item. The matrix of observations need not be complete. An analogous matrix is prepared to describe the relative abilities of the persons, by considering them two at a time and looking only at those items which one got right and the other got wrong. Maximum likelihood estimation and tests for fit of the parameter estimates are examined. (CM)

A FULLY CONDITIONAL ESTIMATION PROCEDURE FOR RASCH MODEL PARAMETERS

Bruce Choppin

CSE Report No. 196
1983

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Eduction
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

The Nature of the Problem

Rasch (1960), in a book on stochastic item response models, set out a "structural model for test items" which subsequently came to bear his name. In this book, Rasch discussed the model's basic assumptions in some detail and began to explore its mathematics. Updating his notation somewhat to conform with current usage, the Rasch model can more simply be written:

$$\text{Probability}\left[X_{vi} = 1\right] = \frac{e^{(\alpha_v - \delta_i)}}{1 + e^{(\alpha_v - \delta_i)}}$$

where X_{vi}, the outcome of person v attempting item i, is one if the response is correct, zero otherwise. α_v is a parameter describing the ability of person v and δ_i is a parameter describing the difficulty of item i.

In most applications, use of the model involves using a large number of observed X values, often arranged in a persons-by-items matrix, to estimate the values of α for the set of people being tested, and δ for the items in the test. In the 1960 book, Rasch makes only hesitant steps towards procedures for estimating α and δ because of limitations in the computational facilities available

to him, and his own preference for simple, graphical methods and intuition. However, he did sketch out an analytic procedure (p. 178-181) for obtaining maximum likelihood estimates of both the α's and the δ's. Unfortunately, this procedure depended on "a mastery of the coefficients $\left\{\begin{matrix} B \\ C \end{matrix}\right\}$ that is not yet at the disposal of the author" in which the persons by items matrix is analyzed. These coefficients represent the number of different but possible patterns of ones and zeros in the matrix that would yield the observed marginal values. Rasch offered some formulae for calculating the coefficients based on summing elementary symmetric functions, and the method was successfully demonstrated by Wright as early as 1965. However, the number of calculations required to determine the elementary symmetric functions increases as k^4 where k is the number of items in the test. Wright points out that this makes the method prohibitively expensive even with the speed and capacity of modern computers, and also inaccurate because round-off errors accumulate during the calculations. In practice this estimation procedure was limited to tests of not more than about ten items. More recently Gustafsson (1977) has reprogrammed the algorithm such that it can handle up to 60 or 70 items satisfactorily. It still remains, however, very expensive when compared to other approaches.

From 1969 on, Wright and various associates at Chicago developed a streamlined procedure based only on the marginals of the observation matrix. The X values, based on the item responses, that led to the

marginals were used only for checking the fit of the data to the model. The preferred statistical method was again that of maximum likelihood. While initial estimates of the δ's are held constant, the α's are adjusted to maximize the likelihood function for the marginals. Then the α values are held fixed while the δ's are adjusted. The cycle repeats until convergence is achieved (Wright, Mead, & Bell, 1979). However, it was pointed out that this approach (dubbed UCON) produces biased estimates, because the ability parameters and their errors of estimate have not been conditioned out of the item difficulty calibration (and vice versa) as they had been in the Rasch procedure described above. Wright and Douglas (1977) proposed a simple correction factor which effectively removed most of the bias and the result was a fast and efficient estimation algorithm that could accurately recover the parameters used to generate artificial data. The algorithm yields standard errors for all the parameter estimates as byproducts of the calculation, and lends itself to several tests of fit of the data to the model. For the last decade, it has been the method used in most Rasch scaling exercises throughout the United States and in a number of other countries.

In educational applications this algorithm's main shortcoming is its inability to handle missing data in an appropriate fashion. More specifically, the algorithm requires a complete rectangular persons-by-items matrix in which each element is a one (representing a correct response) or a zero (representing an incorrect response).

In practice this causes problems:

(a) when, as in survey designs based on matrix sampling or in the use of an item bank, it is deliberately planned that different students should attempt different items,

(b) when the intention is to collect a complete set of data, but for various reasons (e.g., incorrectly assembled tests, student illness, errors in coding or data processing) gaps occur in the set of data prepared for statistical analysis,

(c) when the collected data set is complete but it is desired to edit it selectively (e.g., to remove obvious guesses) before the analysis is carried out.

This paper is directed towards a strategy for overcoming these problems.

The Separation of Ability and Difficulty Parameters.

In his 1960 text, Rasch also described a method of estimation based on the comparison of two or three items at a time (p. 171-174). He gave credit for discovery of the algorithm to G. Leunbach, the Head of the Statistical Unit in the Danish Institute for Educational Research. The main thrust of this algorithm is the manipulation of the data matrix in order to separate out the information needed for the estimation of the item difficulty parameters δ. Conditioning out the ability parameters α in this way avoids the biasing of the parameters described above. In fact, this procedure corresponds closely to conventional practice in the natural sciences: the calibration of instruments, independent of the objects to which they are eventually to be applied, precedes their use for measurement.

The algebraic presentation of this "pairwise" algorithm has been updated to conform to current notation, but it follows the logical sequence used by Rasch.

The basic model we shall use is

$$\text{Prob}\left[X_{vi} = 1\right] = \frac{e^{(\alpha_v - \delta_i)}}{1 + e^{(\alpha_v - \delta_i)}}$$

where X, α_v and δ_i are as defined on page 1.

For many purposes it is simpler to rewrite equation 1 as

$$\text{Prob}\left[X_{vi} = 1\right] = \frac{e^{\alpha_v}}{e^{\alpha_v} + e^{\delta_i}}$$

$$\text{Prob}\left[X_{vi} = 0\right] = \frac{e^{\delta_i}}{e^{\alpha_v} + e^{\delta_i}}$$

which leads to the "odds" (i.e., P/(1-P)) of a correct response

$$\text{Odds}\left[X_{vi} = 1\right] = e^{(\alpha_v - \delta_i)}$$

Consider now the possible outcomes when person v attempts two items i and j. Note that the local independence assumption of the Rasch model requires the responses to the two items be independent. Four separate cases need to considered.

Case (i) - both items correct

$$\text{Prob}\left[a_{vi} = 1,\ a_{vj} = 1\right] = \frac{e^{\alpha_v}}{e^{\alpha_v} + e^{\delta_i}} \cdot \frac{e^{\alpha_v}}{e^{\alpha_v} + e^{\delta_j}}$$

Case (ii) - both items incorrect

$$\text{Prob}\left[a_{vi} = 0,\ a_{vj} = 0\right] = \frac{e^{\delta_i}}{e^{\alpha_v} + e^{\delta_i}} \cdot \frac{e^{\delta_j}}{e^{\alpha_v} + e^{\delta_j}}$$

Case (iii) - item i correct; item j incorrect

$$\text{Prob}\left[a_{vi} = 1,\ a_{vj} = 0\right] = \frac{e^{\alpha_v}}{e^{\alpha_v} + e^{\delta_i}} \cdot \frac{e^{\delta_j}}{e^{\alpha_v} + e^{\delta_j}}$$

Case (iv) - item i incorrect; item j correct

$$\text{Prob}\left[a_{vi} = 0,\ a_{vj} = 1\right] = \frac{e^{\delta_i}}{e^{\alpha_v} + e^{\delta_i}} \cdot \frac{e^{\alpha_v}}{e^{\alpha_v} + e^{\delta_j}}$$

The first two cases hold little interest. A moment's reflection reveals that the information they provide about the ability of person v is distinctly limited, and they provide no information at all about the relative difficulties of items i and j.

Cases (iii) and (iv) are somewhat different. If attention is restricted to these two, for both of which

$$a_{vi} + a_{vj} = 1$$

then we can write

$$\text{Prob}\left[a_{vi}+a_{vj}=1\right] = \frac{e^{(\alpha_v+\delta_j)}}{(e^{\alpha_v}+e^{\delta_i})(e^{\alpha_v}+e^{\delta_j})} + \frac{e^{(\alpha_v+\delta_i)}}{(e^{\alpha_v}+e^{\delta_i})(e^{\alpha_v}+e^{\delta_j})}$$

$$= \frac{e^{\alpha_v}(e^{\delta_i}+e^{\delta_j})}{(e^{\alpha_v}+e^{\delta_i})(e^{\alpha_v}+e^{\delta_j})}$$

If, therefore, we know that person v scored exactly one on the item pair (i, j), then we can write conditional probabilities:

$$\text{Prob}\left[a_{vj}=1 \mid a_{vi}+a_{vj}=1\right] = \frac{\dfrac{e^{(\alpha_v+\delta_i)}}{(e^{\alpha_v}+e^{\delta_i})(e^{\alpha_v}+e^{\delta_j})}}{\dfrac{e^{\alpha_v}(e^{\delta_i}+e^{\delta_j})}{(e^{\alpha_v}+e^{\delta_i})(e^{\alpha_v}+e^{\delta_j})}}$$

$$= \frac{e^{\delta_j}}{(e^{\delta_i}+e^{\delta_j})}$$

and similarly

$$\text{Prob}\left[a_{vj}=1 \mid a_{vi}+a_{vj}=1\right] = \frac{e^{\delta_i}}{(e^{\delta_r}+e^{\delta_i})}$$

The ability parameter α_v has been eliminated entirely from these two expressions. If we know that an individual scores just one on any item pair, the probability that it was one rather than the other that was answered correctly depends solely on the relative difficulty of the two items.

The fundamental importance of this separation of the α and δ parameter sets (by means of conditional probability) to the whole process of measurement has been eloquently described by Rasch (1977). Even when the method is not used in parameter estimation, it is the fact that the model permits the separation that qualifies it for membership in the class of "specifically objective" measurement models--meeting the criteria drawn up by Thurstone as long ago as 1928.

The probability of having a correct response to item i, given that of the two responses to items i and j one is right and one is wrong, can be estimated by observing the results of a large number of people who attempt these two items. If we define b_{ij} to be the number of people who respond correctly to i and incorrectly to j, (with b_{ji} similarly defined) then we can write,

$$\frac{b_{ij}}{b_{ij}+b_{ji}} \text{ as an estimate of } \text{Prob}\left[a_{vi}=1 \mid a_{vi}+a_{vj}=1\right]$$

since this conditional probability does not depend on the α_v and is the same for all people in the group.

i.e., $\frac{b_{ij}}{b_{ij} + b_{ji}}$ estimates $\frac{e^{\delta_j}}{e^{\delta_i} + e^{\delta_j}}$

or $\frac{b_{ij}}{b_{ji}}$ estimates $e^{(\delta_j - \delta_i)}$

which is to say that

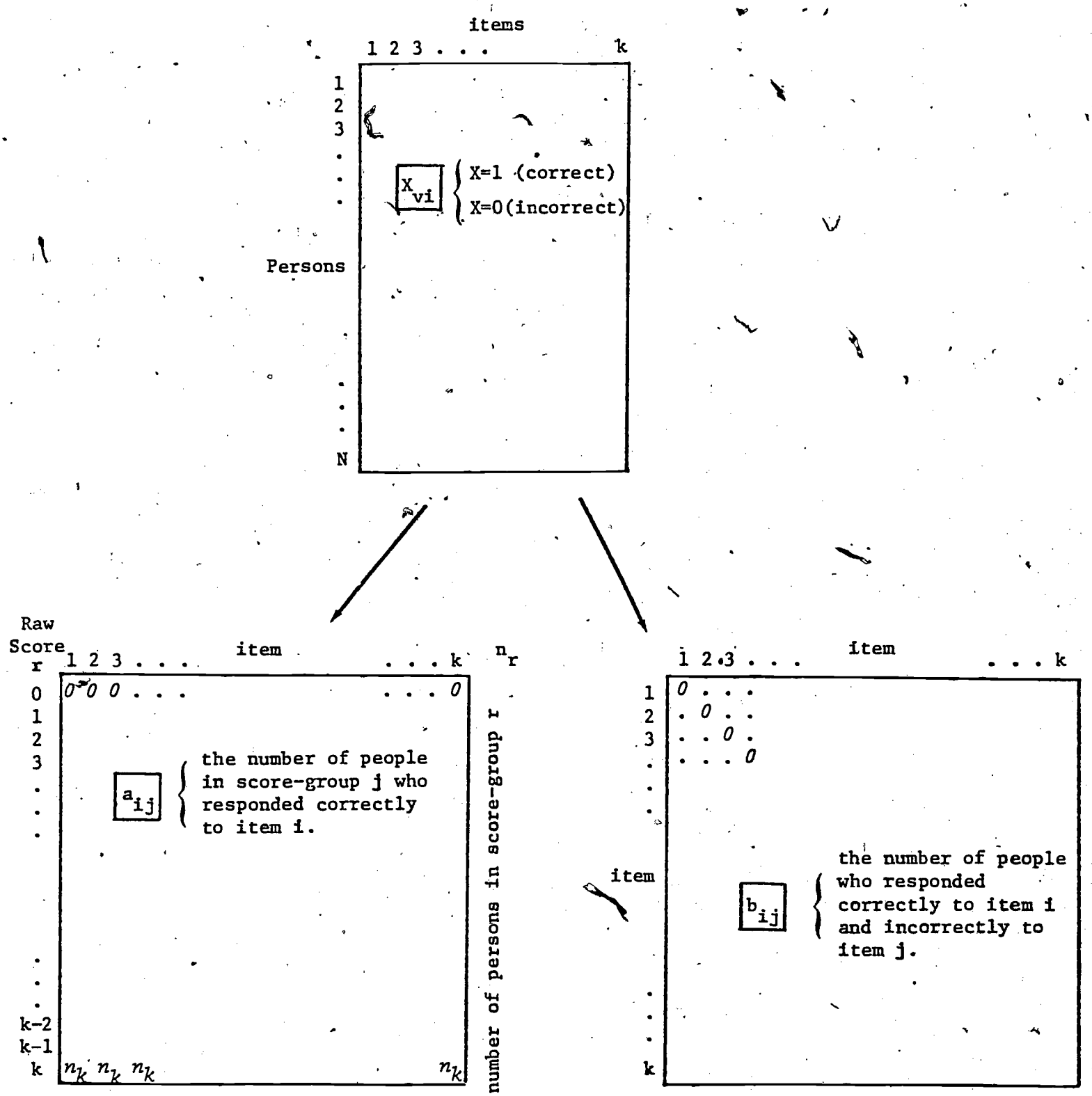
$(\delta_j - \delta_i)$ is estimated by $\log b_{ij} - \log b_{ji}$.

For every pair of items in a test, we can calculate the values of b_{ij} and b_{ji} and hence obtain an estimate of the relative difficulty of the two items concerned. This is more than sufficient information to estimate the relative difficulties of all the items.

The UCON approach described at the beginning makes progress by summarizing the original matrix of 1's and 0's into a (k+1) by k matrix, where for each of (k+1) possible raw scores, the number of correct responses to each of k items is recorded. By contrast, the PAIR approach works by summarizing the original matrix into a k by k matrix of b_{ij} values. However, as can be seen in Figure 1, each summary matrix contains only k(k-1) useful values for the estimation procedure since two rows in the UCON summary matrix have fixed values, and the leading diagonal entries of the PAIR summary are always empty.

Figure 1:

Data Reduction Strategies for Rasch Parameter Estimation



P_i = # correct responses to item i.

1a : Score-group Summarization

1b : Pair-wise Item Summarization

There is of course an analogous matrix of counts describing the relative abilities of the persons, by considering them two at a time and looking only at those items which one got right and the other got wrong. This effectively eliminates δ from the data, and produces a summary matrix with information about the persons. In practice this has been little explored for two reasons. First, there are typically more persons than items in a set of test data, so the person-person summary produces a larger matrix with smaller cell entries. Second, since the ultimate goal is usually to measure the persons individually in terms of their performance on the items, it seems logical first to process the data in order to obtain the best possible calibration of the items, and then to apply these calibrations to the measurement of the persons. Nevertheless the comparison of the measurements obtained by the different methods holds considerable theoretical interest and deserves detailed investigation in the near future. In this paper, however, I shall concentrate on the prior calibration of the items before any measurement of persons is attempted.

Estimating the Difficulty Parameters

To calibrate a set of items from a matrix of observations a complete matrix B is constructed with elements b_{ij} as defined above.

Note that the matrix of observations need not be complete. An individual who is exposed to items i and j gets an opportunity to contribute to b_{ij} or b_{ji}, and thus to the estimation of δ_i and δ_j. It is not necessary for this individual to attempt all (or indeed any) of the other items in the set. This is the algorithm's great strength in practical applications.

The practical solution to the estimation task of item parameters with the pairwise approach, described by Rasch (1960) and first demonstrated by Choppin (1965) at a meeting of the Midwestern Psychological Association, amounted to taking the logarithms of the off-diagonal elements in the B matrix and summing them to get row and column marginals. The difference of the sum of the ith row and the ith column is

$$G_i = \log(b_{1i}) + \log(b_{2i}) + \log(b_{3i}) + \ldots + \log(b_{ki}) - \left[\log(b_{i1}) + \log(b_{i2}) + \log(b_{i3}) + \ldots + \log(b_{ik})\right]$$

$$= \sum_{j \neq i,k}^{j \neq i} \log\left(\frac{b_{ji}}{b_{ij}}\right)$$

which estimates

$$\sum_{j \neq i,k}^{j \neq i} (\delta_i - \delta_j) \quad \text{or} \quad k\delta_i - D$$

where D is the sum of δ_j over <u>all</u> j.

Note that the model, and equation 7 which we have derived from it, has nothing to say about the absolute value of the parameters, only their relative magnitude. If we have a set of α's and δ's which satisfy equation 1, then the new sets produced by adding a constant to all the old values will also satisfy the equation. No "absolute" zero is

defined on the ability or difficulty scale, and it has become conventional to arbitrarily fix the mean δ-value for a particular set of items at 0, since this simplifies the algebra. Parameters can later be adjusted to other zero points, and indeed other units, if so desired.

But taking D = 0, we have the estimation equation

$$G_i = k\delta_i$$

$$\text{or} \quad \delta_i = \frac{G_i}{k}$$

This approach, which is simple and effective when the values of b_{ij} are large and fairly homogeneous, breaks down completely when one or more values of b_{ij} $(i \neq j)$ are zero, since the logarithmic function is not then defined. This appeared to be a major stumbling block, since zero values in the B matrix are met quite often in practice, and led to the approach almost being abandoned.

However, Choppin (1982) pointed out that Rasch's discussion of item triplets (Rasch, 1960, p. 173) can be extended.

If $e^{(\delta_i - \delta_j)}$ can be estimated by $\frac{b_{ji}}{b_{ij}}$, then it can also be estimated by $\frac{b_{jk}}{b_{kj}} \cdot \frac{b_{ki}}{b_{ik}}$ $\left\{\approx e^{(\delta_k - \delta_j + \delta_i - \delta_k)}\right\}$

A better estimate yet is obtained by pooling information across items (i.e., by summing over the double subscript in both top and bottom of the expression). This gives

$$e^{(\delta_i - \delta_j)} \approx \frac{\sum^{k} b_{jk} \cdot b_{ki}}{\sum^{k} b_{kj} \cdot b_{ik}} = \frac{b^*_{ji}}{b^*_{ij}}$$

where b^*_{ij} are the elements of B* the square of the original summary matrix B.

In general B* will not contain off-diagonal zero elements, unless the items are inadequately linked in the sample design (when the complete simultaneous estimation of their difficulties is impossible). Squaring the original B matrix is thus a way of avoiding the problem of zero entries, and leads quickly to a set of δ estimates.

A major drawback, however, is that the manipulation demonstrated in the preceding two paragraphs is only valid for data sets that "fit" the model. In practice some misfit may be expected to occur, and it is important to know about it. The method outlined above will produce δ estimates from virtually any set of test data, and experience suggests that the B* matrix is closer to the stucture prescribed by the model than the original B.

In general, this approach to parameter estimation (which corresponds to a least squares procedure) is not recommended except where strong a priori evidence suggests that data will conform well to the requirements of the model.

Maximum Likelihood Estimation.

A more satisfactory method of unravelling the information stored in the B matrix is that of maximum likelihood. Suppose that in matrix

B, n_{ij} individuals score exactly one on the item pair (i, j), so that we can write:

$$n_{ij} = b_{ij} + b_{ji}$$

Now for any individual in this group, the probability that he gets item i right and item j wrong is:

$$\frac{e^{\delta_j}}{e^{\delta_i}+e^{\delta_j}}$$

and conversely, the probability that i is wrong and j is right is:

$$\frac{e^{\delta_i}}{e^{\delta_i}+e^{\delta_j}}$$

From the binomial theorem, the probability of the n_{ij} individuals dividing into exactly b_{ij} and b_{ji} subgroups is:

$$\frac{n_{ij}!}{b_{ij}!\, b_{ji}!} \cdot \frac{e^{(b_{ij}\,\delta_j + b_{ji}\,\delta_i)}}{(e^{\delta_i}+e^{\delta_j})^{n_{ij}}}$$

and the likelihood of the entire B matrix, given a matrix N of n_{ij} elements is:

$$P[B|N] = \prod^{i,j,i\neq j} \left(\frac{n_{ij}!}{b_{ij}!\, b_{ji}!} \cdot \frac{e^{(b_{ij}\delta_j + b_{ji}\delta_i)}}{(e^{\delta_i} + e^{\delta_j})^{n_{ij}}} \right)$$

The problem now is to find a set of δ_i's which maximize the value of this function. This maximum and the maximum of its logarithm occur at the same point, so for simplicity we may write the log likelihood:

$$L = C + \sum^{i,j,i\neq j} (b_{ij}\delta_j + b_{ji}\delta_i) - \sum^{i,j,i\neq j} (b_{ij} + b_{ji}) \log (e^{\delta_i} + e^{\delta_j})$$

where C is a function of the b 's but not of the δ's.

For maximum likelihood $\frac{\partial L}{\partial \delta_i} = 0$ for each i

$$\text{i.e.} \quad 0 = \frac{\partial L}{\partial \delta_i} = \sum^{j} b_{ji} - \sum^{j} \left(\frac{e^{\delta_i} \cdot (b_{ij}+b_{ji})}{(e^{\delta_i} + e^{\delta_j})} \right) \qquad \text{(all i)}$$

$$\text{or} \quad \sum^{j} b_{ji} = \sum^{j} \left(\frac{(b_{ij} + b_{ji})\, e^{\delta_i}}{(e^{\delta_i} + e^{\delta_j})} \right) \qquad \text{(all i)}$$

As before, it is necessary to insert an additional linear constraint on the set of δ's since it is clear that adding any constant value to all the δ's in the above equations would not change the nature of the solution. Rasch scaling deals in relative difficulty rather than absolute difficulty and no zero point on the scale can be uniquely defined. However, once the δ-value for one item has been fixed (if necessarily arbitrarily) then all the other δ's can be defined by relating them to the first. The usual constraint is to put the sum of the δ-values equal to zero.

This set of k equations in k unknowns can be solved by various iterative techniques. Two that have been found to work well in practice are:

$$^{(n+1)}\delta = \log\left(\sum^{j} b_{ij}\right) - \log\left(\sum^{j} \frac{(b_{ij}+b_{ji})}{(e^{\delta_i}+e^{\delta_j})}\right)$$

and the Newton-Raphson procedure:

$$^{(n+1)}\delta_i = {}^{(n)}\delta_i - \frac{\sum^{j} b_{ij} - \sum^{j} \frac{(b_{ij}+b_{ji})}{(e^{\delta_i}+e^{\delta_j})} \cdot e^{\delta_i}}{\sum^{j} \frac{(b_{ij}+b_{ji})e^{(\delta_i+\delta_j)}}{(e^{\delta_i}+e^{\delta_j})}}$$

In general an efficient procedure has proved to be to set the initial $^{(o)}\delta$ - values all equal to zero, and then to apply the first iterative set of equations three or four times. This produces a set of reasonably good approximations which, when used with the Newton-Raphson equations, leads rapidly to convergence and a solution for the various δ's.

A great advantage of the maximum likelihood procedure is that it can be used to generate standard errors for the estimated parameters (Kendall & Stuart, 1969). The variance-covariance matrix V is the inverse of a matrix whose elements are:

$$\left[-\frac{\partial^2 L}{\partial\delta_i \partial\delta_j}\right]$$

evaluated at the maximum likelihood solution. In practice, however, a simpler approximation

$$s_i = \sqrt{\frac{k}{-\frac{\partial^2 L}{\partial\delta_i^2}}}$$

seems adequate, and is recommended for routine use.

Estimating the Ability Parameters

If $\underline{X_{vi}}$ $(i = 1, k)$ are the set of scored responses for person v, whose total test score is r_v, and whose ability parameter is α_v; then we may write:

$$r_v = \sum^{i} X_{vi}$$

where the summation runs over the set of items attempted by person v.

The likelihood of the response X_{vi} according to the model is:

$$\frac{e^{X_{vi}(\alpha_v - \delta_i)}}{1 + e^{(\alpha_v - \delta_i)}}$$

since X_{vi} takes values one and zero accordingly as the response is right or wrong.

From this, the likelihood function for the entire set of responses $(\underline{X_{vi}})$ for person v is:

$$\prod^{i} \frac{e^{X_{vi}(\alpha_v - \delta_i)}}{1 + e^{(\alpha_v - \delta_i)}}$$

The logarithm of this function:

$$L = \left(\sum^{i} X_{vi}\alpha_v - \sum^{i} X_{vi}\delta_i \right) - \sum^{i} \log\left(1 + e^{(\alpha_v - \delta_i)}\right)$$

$$= \left(\alpha_v r_v - \sum^{i} X_{vi}\delta_i \right) - \sum^{i} \log\left(1 + e^{(\alpha_v - \delta_i)}\right)$$

For the ML solution, $\frac{dL}{d\alpha_v} = 0$. It should be noted that in this case the δ's are regarded as already known, so that α_v is the only parameter to be estimated.

$$0 = \frac{dL}{d\alpha_v} = r_v - \sum^{i} \frac{e^{(\alpha_v - \delta_i)}}{1 + e^{(\alpha_v - \delta_i)}}$$

This equation does not contain the item response (X_{vi}). It demonstrates a result, already obtained by other writers, that the ability estimate depends not upon the particular pattern of item responses obtained, but only upon the "total score." r is a sufficient statistic for ability, and the conventional practice of using total scores as measures has a logical foundation.

The equation $$r_v = \sum^{i} \frac{e^{(\alpha_v - \delta_i)}}{1 + e^{(\alpha_v - \delta_i)}}$$

can be solved for α_v.

The score on the test takes values 0, 1, 2, ... k. Each of the k terms on the right hand side of the equation lies between 0 and 1 for real values of X_{vi} and δ_i. Note that there are no solutions for $r = 0$ and $r = k$. For these values the likelihood function has no maximum, and this could have been anticipated. If an individual responds correctly to every item (i.e., $r_v = k$), then we have no information on which to base any upper bound for an ability estimate. Similarly, if every item is answered incorrectly, there are no data to suggest just how low the level of ability might be.

Note that once a set of items has been calibrated (i.e., the δ's have been estimated), it is possible to estimate an ability parameter for each possible score on the test, regardless of whether or not any individual actually obtains such a score. If a test is constructed by selecting items from an already calibrated item bank, then ability parameters for all possible scores on the new test can be calculated even before the test is used.

The standard errors of the ability parameters, corresponding as they do to the standard error of measurement, are usually of more interest than the standard errors of the item difficulties. Furthermore, they are typically considerably larger, since the ability parameter estimates are based upon only k observations (usually between 10 and 100) whereas item calibration is typically based upon the results of at least several hundred individuals.

In general, if we assume that the δ's are established with some precision, the standard errors of the α's can be developed from the

log likelihood function.

$$L = \alpha_v r_v \sum^{i} X_{vi}\delta_i - \sum^{i} \log[1 + e^{(\alpha_v - \delta_i)}]$$

$$\frac{dL}{d\alpha_v} = r_v - \sum^{i} \frac{e^{(\alpha_v - \delta_i)}}{1 + e^{(\alpha_v - \delta_i)}}$$

$$\frac{d^2L}{d\alpha_v^2} = -\sum^{i} \frac{[1 + e^{(\alpha_v - \delta_i)}]e^{(\alpha_v - \delta_i)} - e^{2(\alpha_v - \delta_i)}}{[1 + e^{(\alpha_v - \delta_i)}]^2}$$

$$= -\sum^{i} \frac{e^{(\alpha_v - \delta_i)}}{[1 + e^{(\alpha_v - \delta_i)}]^2}$$

Then the standard error of measurement for an individual who receives an ability estimate α_v for his responses to items with difficulties δ_i is:

$$\sqrt{\frac{-1}{\frac{d^2L}{d\alpha_v^2}}}$$

The second differential reaches a maximum value of $\frac{-k}{4}$ when all

the δ's are equal to α_v. In practice, if the δ's are all fairly close to α_v (i.e., if the items are all closely matched to the person's ability), then the second differential remains close to $-\frac{k}{4}$ and so the standard error is given approximately by $\frac{2}{\sqrt{k}}$ logits; in general, however, we can write

$$s_{\alpha_v} > \frac{2}{\sqrt{k}} \text{ logits}$$

whatever the distribution of the δ's.

Tests of Fit

Control of the model by validating its conformity to the structure of a particular data set, within pre-specifiable limits, has been difficult to achieve with the PAIR method of parameter estimation. The most frequently used approach has been the non-random splitting of the original data set into two parts based on some characteristic of the persons, calibrating the full set of items for each part of the data separately, and plotting the results against one another. This is inexpensive, straightforward, and has considerable utility although it lacks mathematical elegance. A division of the sample of persons into high performers and low performers at the median raw score is the most severe test of the anticipated invariance of item difficulty parameters. It focuses directly on the assumption

of equal discriminating power for all items in the set, and the associated though contrary threat, ability-related random guessing. Since the same set of item parameters is being estimated for both high and low ability groups, and the mean difficulty in each case is being fixed at zero, the model predicts that within the limits of sampling error the same item calibrations should emerge from each half of the analysis. Figure 2 demonstrates typical results from two multiple-choice achievement tests, one of which fits the model very well and one of which shows considerable evidence of guessing. Experience with plots of this type shows that they can be very informative, and Figure 3 shows a somewhat simplified guide to their interpretation.

Of course other splits can be used to generate these plots. For example, to test for the presence of sex bias within a test it is possible to plot calibration obtained from males against those obtained from females. The plot reproduced in Figure 4 contains item difficulties for a mathematics test calibrated for groups of students who studied two different curricula. Analysis of the discrepancies from the predicted straight line showed how the pattern of learning produced by the new curriculum was different from that of the old (and these differences were not in accord with the intentions of the curriculum development team. Choppin, 1977).

A more detailed control of the model requires going back to the original persons-by-items data matrix and estimating the probability of a correct response for each person/item interaction based on the

Figure 2A : Cross-calibration of items for a test that 'fits'.

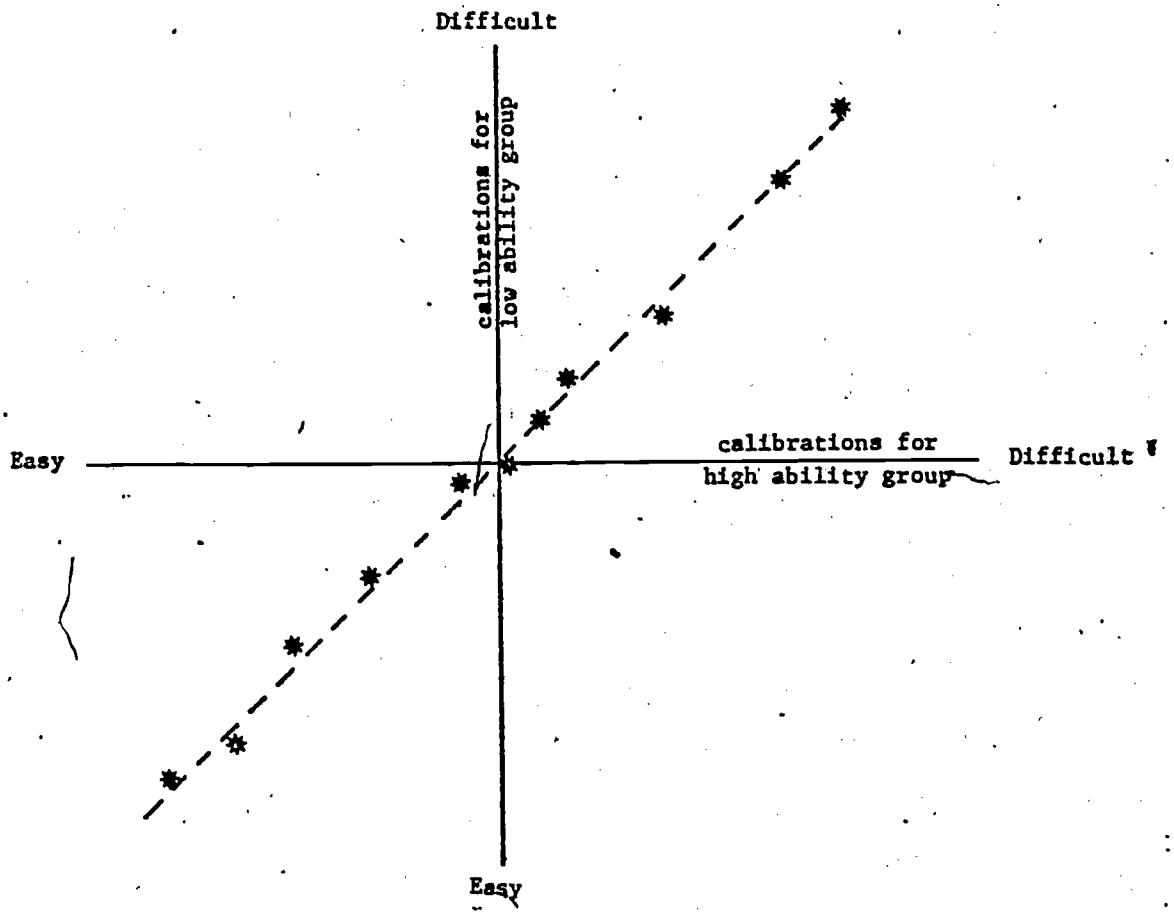


Figure 2B : Cross-calibration of items demonstrating guessing.

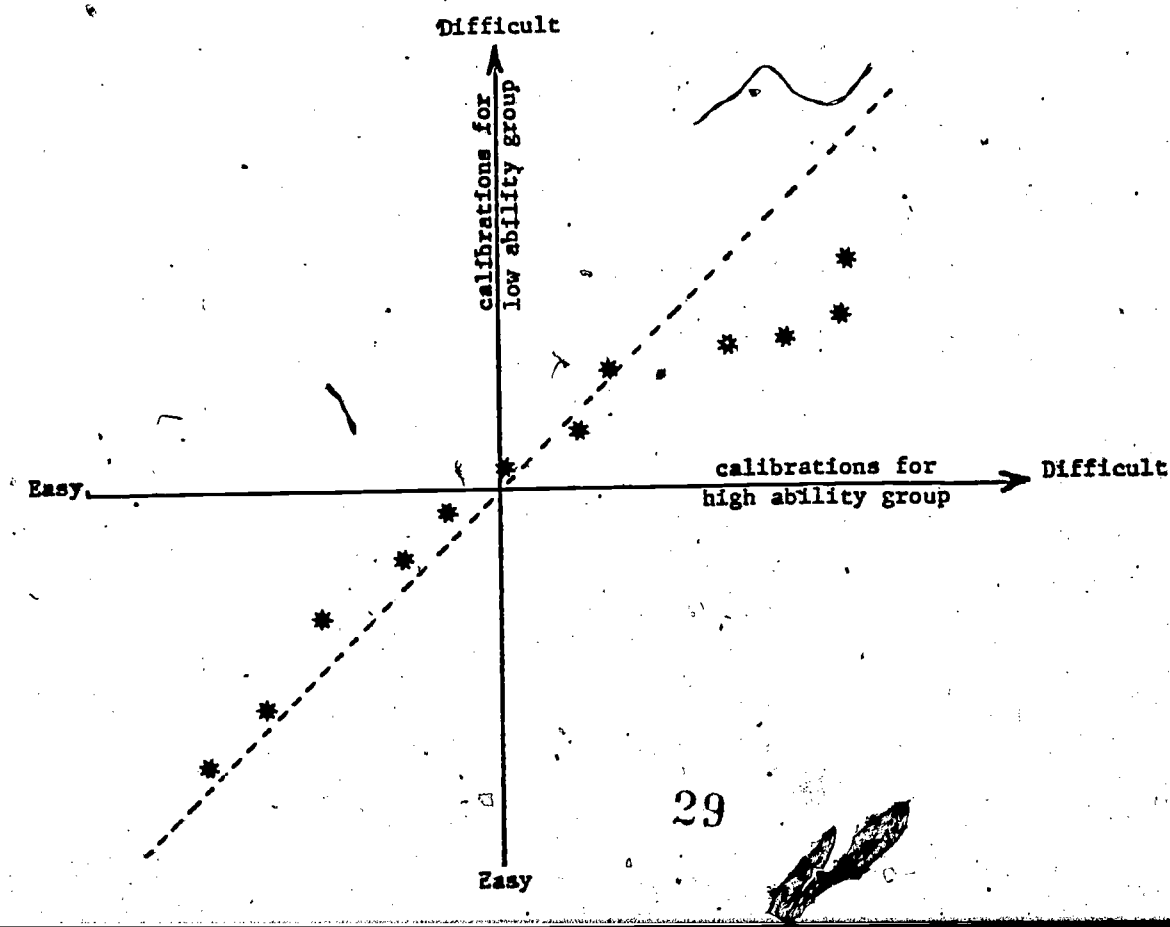


Figure 3: Guide to Interpretation

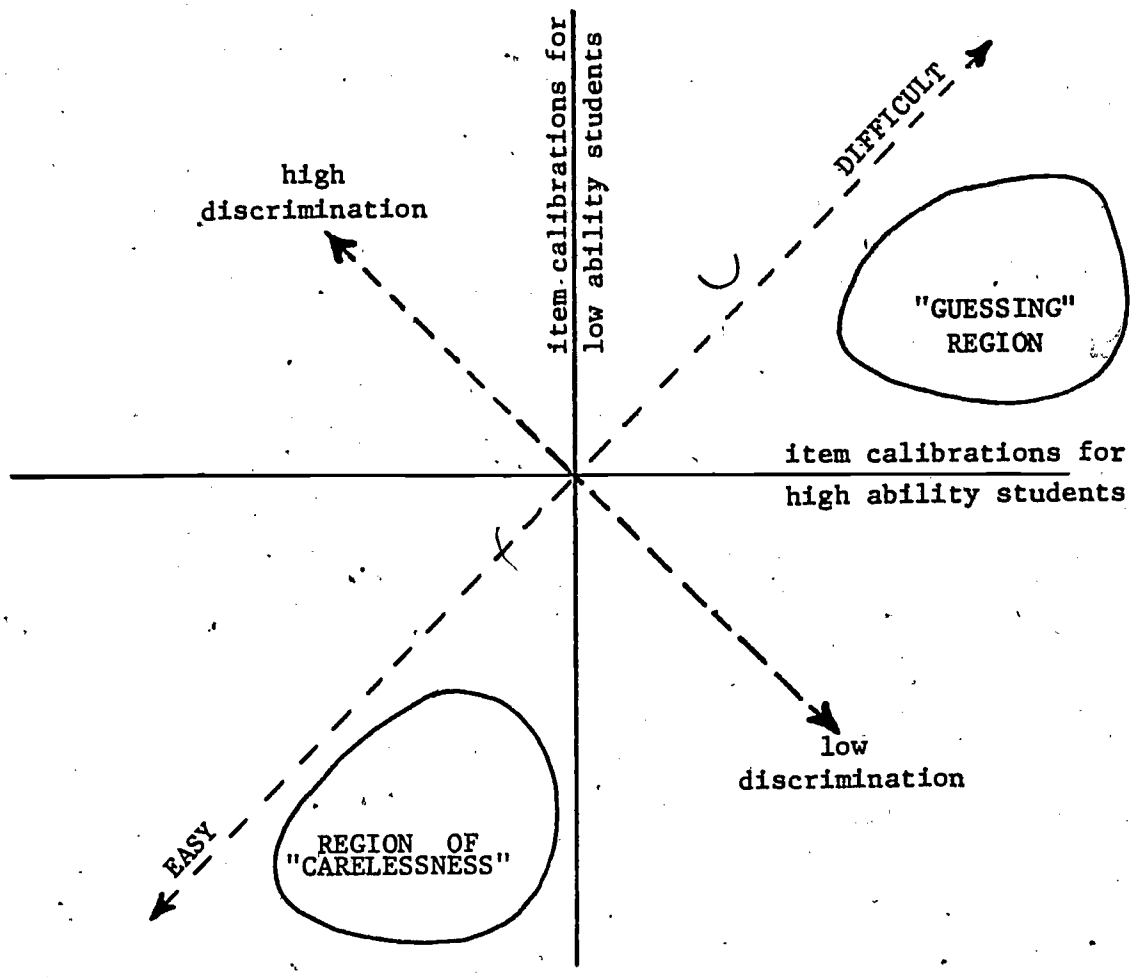
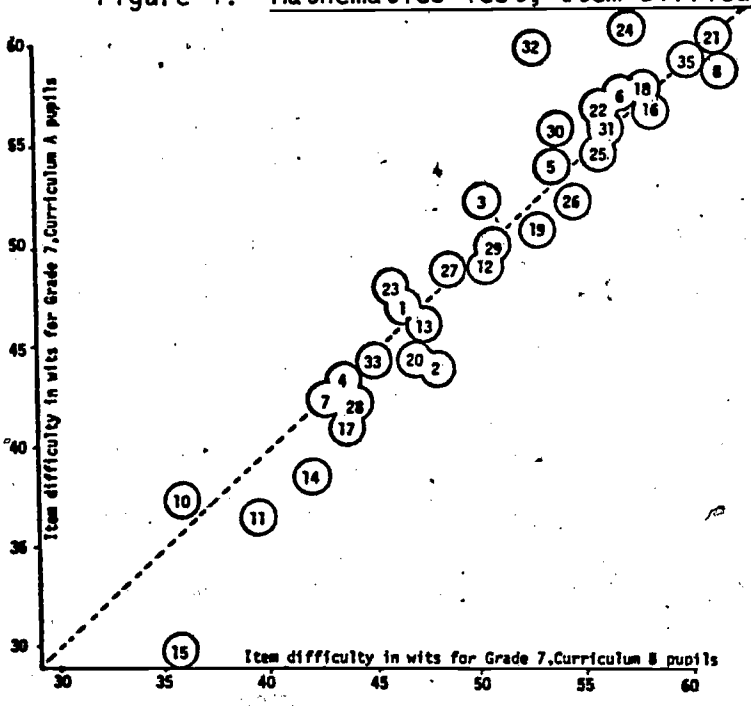


Figure 4: Mathematics Test, Item Difficulty



estimates of α and δ. When these probabilities are compared to the observations, a matrix of residuals is generated. This topic has been well covered in Mead (1975) and Wright and Stone (1979) and will not be further developed here.

Rasch (1960, p. 174) suggested that the examination of item triplets might offer an effective control of the model, but this has not proved to be the case. Recently, however, it has been noticed that the comparison of the B and B* matrices offers a concise test of the local independence assumption. The ratio $\left[\frac{b_{ji}}{b_{ij}}\right]$ estimates $(\delta_i - \delta_j)$ ignoring all other items, whereas $\left[\frac{b^*_{ji}}{b^*_{ij}}\right]$ estimates only through the comparison of i and j to the other items. If there exists a local contextual effect (e.g., if item 15 is easier than it would otherwise be because it comes immediately after item 14), then the b and b* values should show it. This comparison is accomplished by a χ^2 statistic. The method holds considerable promise since the assumption of local independence has been strongly attacked as unrealistic (Goldstein, 1979). A number of studies of achievement test data in which items are administered in different orders and with or without other groups of items suggests that the local independence assumption is often well met in practice, although other evidence (Tang, 1982) suggests that on a test of reasoning skills such as a

progressive matrices test, the context is extremely important. It seems probable that the (B,B*) comparison will be used for testing local independence even when parameter estimation is achieved through UCON or maximum likelihood PAIR.

REFERENCES

Choppin, B.H. Pairwise calibration of item difficulty using the Rasch model. Paper read to meeting of the Midwestern Psychological Association, Chicago, 1965.

Choppin, B.H. An item bank using sample free calibration. Nature, 1968, 219, 870-872.

Choppin, B.H. Comparing curricula by means of Rasch scaling. Invited address Weizmann Institute of Science, Rehovot, Israel, 1977.

Choppin, B.H. Item banking and the monitoring of achievement. Slough, England: National Foundation for Educational Research, 1978.

Choppin, B.H. The Rasch scaling model and application. BP3K; Ministry of Education and Culture, Jakarta, 1982(a).

Choppin, B.H. The use of latent-trait models in the measurement of cognitive abilities and skills. In D. Spearitt (Ed.) The improvement of measurement in education and psychology, Melbourne: ACER, 1982(b).

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

Gustafsson, J.E. The Rasch model for dichotomous items. Research Report 63. Institute of Education, University of Goteberg, 1977.

Kendall, M.G., & Stuart, A. Advanced theory of statistics. Hafner, New York, 1969.

Mead, R.J. Analysis of fit to the Rasch model. Doctoral dissertation, University of Chicago, 1975.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960. (Reprinted by University of Chicago Press, 1980)

Rasch, G. On specific objectivity. Danish Yearbook of Philosophy, 1977, 14, 58-94.

Tang, E.C. Personal communication, 1982.

Thurstone, L.L. The measurement of opinion. Journal of Abnormal and Social Psychology, 1928, 22, 415-430.

Wright, B.D., & Douglas, G.A. Conditional versus unconditional procedures for sample free item analysis. Educational and Psychological Measurement, 1977, 37, 573-586.

Wright, B.D., Mead, R.J., & Bell, S.R. BICAL: Calibrating items with the Rasch model. Research Memorandum 23B, Statistics Lab, Education Department, University of Chicago, 1979.

Wright, B.D., & Stone, M.H. Best test design. Chicago: MESA Press, 1979.