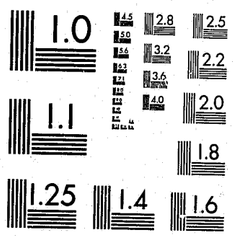

BSSR

REAU OF SOCIAL SCIENCE RESEARCH, INC.

WASHINGTON, D. C.

72749

A REVIEW OF ESTIMATION PROCEDURES FOR THE RASCH MODEL WITH AN EYE TOWARD LONGISH TESTS

By

Howard Wainer and Anne Morgan
The Bureau of Social Science Research, Inc.

and

Jan-Eric Gustafsson
Institute of Education
University of Goteborg

"Longtemps, je me suis couché de bonne heure."
(Proust,1913;p. 3)

ABSTRACT

Two estimation procedures for the Rasch Model are reviewed in detail, particularly with respect to new developments that make the more statistically rigorous Conditional Maximum Likelihood estimation practical for use with longish tests. Emphasis of the review is on European developments which are not well known in the English writing world.

I. INTRODUCTION

Classical test theory models based on the concept of true-score have been in use for most of the 20th century, but for the last decade or so their shortcomings have become increasingly apparent . The principal shortcoming is that with classical test theory the item parameters change with the norming group. The development of latent trait theory has resulted from the search for a replacement. Lord and Novick (1968) honestly reflect this state of affairs by rigorously developing the various true-score models with their strengths and weaknesses and also specifying (primarily in the chapters by Birnbaum) the basis of the emerging latent trait technology. This new technology has been rather slow to catch on for a variety of reasons; among which mathematical and numerical problems in parameter estimation are probably the most important.

There has been, and remains, considerable debate about the best single model to use for the scoring of tests. We shall not enter this debate, but instead will stay within the confines of the simplest model, the one-parameter logistic (usually called "The Rasch Model" after its originator [Rasch, 1960;1966]). If this simple model fits the data, there is no need for the more complex ones. The question of whether the model fits the data or not can be answered, partially at least, with statistical goodness-of-fit tests. Even for this simplest version, however, there are serious problems of parameter estimation. The problems can be simply stated: estimation methods that are statistically rigorous could not be used for the kind of tests most likely to be scored with a latent trait model (large-scale standardized tests with many people taking them [like the SAT] that are often quite long). Short-cut methods and approximations have been devised

which appear to work quite well (from simulations) but are still not statistically rigorous (e.g., Fischer, 1974; Wright & Douglas, 1976; Wright & Mead, 1977).

All at once a variety of developments have occurred which seem to have resolved this problem. These developments have not occurred at any one time, nor at any one place, nor are they by any one person, but they are all here now and can profitably be taken advantage of. This paper reviews these developments and tries to catalog them with respect to how each can be brought to bear on the problem of the estimation of parameters of the Rasch Model for moderate to long tests (40 to 90 items). In this paper we shall report the approximation methods of Wright and his colleagues, the developments of Fischer and Scheiblechner, the numerical break-through of Gustafsson, and the work on tests of fit that Andersen and Martin-Löf have done. Much of this material is not in the English language literature.

The Rasch Model is a latent trait model of a very simple nature: the probability of a correct answer to an item is a function of the difficulty of that item and the ability of the person.

The model makes the following assumptions:

(1) All items measure the same trait. The test is then called homogeneous.

(2) The item characteristic curve, the function relating the probability of a correct answer to an item to the underlying ability variable (the latent trait), has a logistic form.

(3) Local stochastic independence of the items (i.e., whether or not a person solves an item depends on that person's ability and on the difficulty of the items, but not on which other items she or he has previously solved).

If these assumptions hold, the following properties are obtained as well:

(a) The raw (number correct) score is a sufficient statistic for the estimation of ability.

(b) The comparison of two people is fully described by the difference in their abilities on the latent ability dimension. This does not depend on which specific items were administered to them (item independent person measurement).

(c) The estimation of item difficulties is independent of the ability of the sample on which they were calibrated (sample-free item calibration).

These last two properties of the model are very important. In 1950 Gulliksen characterized current thinking when he wrote, "A significant contribution to item analysis theory would be the discovery of item parameters that remain relatively stable as the item analysis group changed" (p. 392). The Rasch model and latent trait based models in general satisfy Gulliksen's requirement that the results of a person do not depend either on which reference population that person belongs to, nor on the selection of a specific set of items from the homogeneous universe of items, if the data fit the model. This emphasizes the great potential importance of the Rasch Model to ability testing, should it be found that test data fit the model reasonably well.

II. THE RASCH MODEL FOR DICHOTOMOUS DATA

The response of person v to item i is denoted A_{vi}, it can take values 0 (incorrect) or 1 (correct). The probability of a correct response according to the Rasch Model is given by

$$P(A_{vi} = 1 \mid \xi_v, \delta_i) = \frac{\exp(\xi_v - \delta_i)}{1 + \exp(\xi_v - \delta_i)} \qquad (2.1)$$

where δ_i, $i = 1,\ldots,k$, is the item parameter describing the difficulty of item i and ξ_v, $v = 1,\ldots,n$, is the ability parameter describing the ability of person v. Both of these parameters are in the logistic metric and are referred to as measurement in "logits." An alternative, and frequently useful, representation involves an exponential transformation yielding $\theta_v = \exp(\xi_v)$ and $\varepsilon_i = \exp[-\delta_i]$. Using this change in variable allows one to rewrite equation (2.1) as

$$P(A_{vi} = 1 \mid \theta_v, \varepsilon_i) = \frac{\theta_v \varepsilon_i}{1 + \theta_v \varepsilon_i} \qquad (2.2)$$

where ε_i is scaled in the opposite direction of δ_i, and is usually interpreted as the "easiness" of item i.

The probability of the response a_{vi} (a more general case of equation [2.2]) can be written

$$P(A_{vi} = a_{vi} \mid \theta_v, \varepsilon_i) = \frac{(\theta_v \varepsilon_i)^{a_{vi}}}{1 + \theta_v \varepsilon_i} \qquad (2.3)$$

or similarly

$$P(A_{vi} = a_{vi} \mid \xi_v, \delta_i) = \frac{\exp[a_{vi}(\xi_v - \delta_i)]}{1 + \exp(\xi_v - \delta_i)} \qquad (2.4)$$

Both of these representations (2.3 and 2.4) will be used in the subsequent elaboration of the Rasch model.

III. ESTIMATION PROCEDURES WITH RESPECT TO LONGISH TESTS

In this section of the paper we shall describe two estimations procedures[1]. These are the Unconditional Maximum Likelihood[2] (UML) and the Conditional Maximum Likelihood (CML) methods. Until recently, only the UML could be practically applied for longish tests (those with more than 30 or 40 items). This has changed recently with newer and more sophisticated estimation schemes, better numerical methods, and faster computers. In the past there were strong theoretical reasons for preferring the CML method, but it has not been feasible to apply it to longish tests. Wright and his colleagues have corrected some of the difficulties of the

[1] Two less commonly known estimation procedures, which have been used for longish tests, should be mentioned (see Fischer, 1970, 1974 for details): a "minimum-chi-square method" (Fischer, 1970), a very fast algorithm with consistent estimators but for which the mathematical statistical basis is incomplete: Scheiblechner's (1971) conditional maximum likelihood algorithm, whose computer program can be used for a maximum of 50 items.

[2] The term "Unconditional Maximum Likelihood" estimation is an unfortunate one in this application, since this is actually "JOINT Maximum Likelihood" estimation. Unconditional estimation is when a part of the parameter space is integrated out by assuming a distribution and integrating over it. Joint estimation is when estimates are obtained for all parameters simultaneously by maximizing the likelihood in all directions at once. Nevertheless, to maintain congruence with current usage, and so avoid confusion, we shall use the term Unconditional.

UML, making it an acceptable method as far as bias is concerned (Wright & Douglas, 1977b); Fischer, Gustafsson, Martin-Löf and others have advanced CML to the point where it is practical for longish tests. We shall present both methods and will comment on the choice between them in a later section.

The Unconditional Maximum Likelihood Estimation Procedure (UML)

This method of estimating the item and ability parameters of the Rasch Model simultaneously was presented by Wright and Panchapakesan(1969) and Fischer and Scheiblechner (1970). It yields a solution that must be corrected for bias, but until recently it was the only viable method for tests over 30-40 items.

The basic data matrix from which estimation proceeds is the matrix A having elements $\{a_{vi}\}$, which is, say persons by items. Summing across items yields the raw score for person v, denoted r_v. Summing across persons yields the total number of correct responses to item i, and is denoted s_i. Under the assumption of local stochastic independence the likelihood of A is the product of the probabilities of all the entries of A. This is denoted by Λ and is shown in equation (3.1):

$$\Lambda = \prod_{v=1}^{n} \prod_{i=1}^{k} \frac{(\theta_v \varepsilon_i)^{a_{vi}}}{1+\theta_v \varepsilon_i} = \frac{\prod_v \theta_v^{r_v} \prod_i \varepsilon_i^{s_i}}{\prod_v \prod_i (1+\theta_v \varepsilon_i)} \qquad (3.1)$$

From this likelihood function, it is immediately apparent that only the marginal sums of A are represented [r_v and s_i]. Thus, one need not take into account the "inner structure" of A as yet, that is, which items a certain examinee has answered, nor which examinees answered a particular item (this will enter in later on when the goodness-of-fit of the model is examined). Raw score (number correct) is a sufficient statistic for estimating person parameters and item score is a sufficient statistic for item parameters.

The likelihood function can be maximized in the usual way by first noting that the log of the likelihood function achieves its maximum at the same place as the function, and then using the multivariate form of Newton-Raphson on the log likelihood function. This is accomplished by calculating the gradient vectors with respect to each parameter, and the Hessian (matrix of second partials).

The details of the derivation of the estimation equations is found in Wright's work (e.g. Wright & Mead, 1977; Wright & Stone, 1979). There is an indeterminacy in the model which can be removed by imposing some sort of normalization. One way of normalizing is to set an origin (say the difficulty of an item equals zero), or set a scale (say the sum of all difficulties equals unity).

That maximum likelihood estimates are not consistent in certain situations has been known since 1948 (Neyman & Scott). This is the case for the Rasch Model when structural parameters (the item difficulties) are estimated in the presence of incidental parameters (the person abilities). When sample size is increased the problem, of course, remains since each new person brings a new incidental parameter. If, however, the estimation equating can be formulated in the item parameters only, consistency and unbiasedness is assured (Andersen, 1977). This can be done if there exists a minimal sufficient statistic for the person parameters, and in the Rasch model, raw score is such an estimator.

Wright and Douglas(1977b) have shown bias to average (over scores and items) k/(k-1). Applying this correction factor to the difficulty estimates allows one to obtain estimates that are, on average, unbiased and in simulations seem to be similar to those obtained by the unbiased CML method to be discussed next.

The following is an algorithm for the unconditional estimation of item and person parameters from Wright and Douglas (1977b):

(1) Calculate s_i, the total number of correct responses to item i and, n_r the number of persons with raw score r.

(2) Edit the data to exclude zero or perfect scores for both items and persons (i.e. r_v = 0 or k and s_i = 0 or n).

(3) Initialize a starting vector $\underline{b}_r$ as,

$$b_r^0 = \log[r/(k-r)] \quad \text{for } r = 1,\ldots,k-1. \tag{3.2}$$

(4) Initialize a vector $\underline{d}_i$, centered at d. = 0 as,

$$d_i^0 = \log\left[\frac{N - s_i}{s_i}\right] - \sum_i^k \log\left[\frac{N - s_i}{s_i}\right]/k \qquad i = 1,\ldots,k \tag{3.3}$$

(5) Improve each estimate $\underline{d}_i$ by applying equation 3.4

$$d_i^{j+1} = d_i^j - \left[\frac{-s_i + \sum_r^{k-1} n_r p_{ri}^j}{-\sum_r^{k-1} n_r p_{ri}^j \ (1 - p_{ri}^j)}\right] \qquad (i=1,\ldots,k) \tag{3.4}$$

until convergence at some reasonable criterion, say CRIT.

$$|d_i^{j+1} - d_i^j| < \text{CRIT},$$

where CRIT = .01 is a good value, and

$$p_{ri}^j = [\exp\{b_r - d_i^j\}]/[1 + \exp\{b_r - d_i^j\}]$$

(6) Recenter the vector $\underline{d}_i$ at d. = 0.

(7) Using the improved vector $\underline{d}_i$, apply equation 3.5 to improve each b_r.

$$b_r^{m+1} = b_r^m - \left[\frac{r - \sum_i^k p_{ri}^m}{-\sum_i^k p_{ri}^m \ (1 - p_{ri}^m)}\right] \qquad (r=1,\ldots,k-1) \tag{3.5}$$

until convergence at

$$|b_r^{m+1} - b_r^m| < \text{CRIT}$$

where $P_{ri}^m = [\exp\{b_r^m - d_i\}]/[1+\exp\{b_r^m - d_i\}]$

(8) Repeat steps (5) through (7) until successive estimates of $\underline{d}_i$ become stable, that is,

$$\sum_i [d_i^{\ell+1} - d_i^{\ell}]^2 / k < (\text{CRIT})^2$$

(9) Correct for bias by multiplying each d_i by (k-1)/k

(10) Calculate the $\underline{b}_r$ for those corrected $\underline{d}_i$.

(11) Correct the bias by multiplying each b_r by (k-2)/(k-1)

(12) Calculate the asymptotic estimates of the standard errors of difficulty estimate from the inverse

Hessian (equation 3.6),

$$SE[d_i] = \sum_r \{n_r P_{ri}[1-P_{ri}]\}^{-1/2}. \qquad (3.6)$$

When the test score distribution is symmetric and the tests are rather long, a very economical procedure for approximating the UML method was devised by Cohen (1979)

The Conditional Maximum Likelihood Procedure (CML)

The following description of the conditional approach follows closely that given by Gustafsson (1977).

Consider a given examinee with the raw score r_v, corresponding to the person parameter θ_v. The probability of obtaining any raw score vector $(\underline{a}_{vi})$ given the person parameter and the vector of item parameters is:

$$P\{(a_{vi})|\theta_v,(\varepsilon_i)\} = \prod_{i=1}^{k} \frac{(\theta_v\varepsilon_i)^{a_{vi}}}{1+\theta_v\varepsilon_i} = \frac{\theta_v^{r_v}\prod_i \varepsilon_i^{a_{vi}}}{\prod_i (1+\theta_v\varepsilon_i)} \qquad (3.7)$$

To be able to express this probability as a conditional probability of a given score r_v, one must first know the probability of obtaining score r_v given θ_v. This latter probability is given by the sum of the probabilities of all possible ways of obtaining the score r_v, that is, the sum of all terms described in (3.6) in which the vector a_{vi} sums to r.

Any given score r on a set of k items can be obtained in $\binom{k}{r}$ different ways. A special notation is needed to express this in a simple way. Define the elementary symmetric function of order r in the parameters ε_i as:

$$\gamma_r\{(\varepsilon_i)\} = \sum_{\Sigma a_{vi}=r} \prod_{i=1}^{k} \varepsilon_i^{a_{vi}} \qquad (3.8)$$

In the expansion of this sum of products, the summation is made over those $\binom{k}{r}$ combinations in which $\sum_i a_{vi} = r$. For simplicity, γ_r will be used to denote the symmetric function of order r in the item parameters.

This new notation allows us to write the probability of obtaining the score r given θ_v and ε_i as:

$$P\{r|\theta_v,(\varepsilon_i)\} = \sum_{\Sigma a_{vi}=r} \prod_{i=1}^{k} \frac{(\theta_v\varepsilon_i)^{a_{vi}}}{1+\theta_v\varepsilon_i} = \frac{\theta_v^{r_v}\gamma_r}{\prod_i (1+\theta_v\varepsilon_i)} \qquad (3.9)$$

The conditional probability of obtaining the vector $\underline{a}_{vi}$ with the total score r_v, given the score r_v is thus given by equations 3.7 through 3.9, yielding:

$$P\{(a_{vi})|r,(\varepsilon_i)\} = \frac{P\{(a_{vi})|\theta_v,(\varepsilon_i)\}}{P\{r|\theta_v,(\varepsilon_i)\}} = \frac{\prod_{i=1}^{k} \varepsilon_i^{a_{vi}}}{\gamma_r} \qquad (3.10)$$

Note that this conditional probability is not a function of θ_v, but only of the item parameters.

Using the assumption of conditional independence one can now obtain the conditional likelihood of the data matrix A, with elements $\{a_{vi}\}$ for n persons yielding:

$$A = \prod_{v=1}^{n} \frac{\prod_{i=1}^{k} \epsilon_i^{a_{vi}}}{\gamma_{r_v}} \qquad (3.11)$$

Denoting n_r as the number of persons with raw score r (1,...,k-1) and s_i as the score of item i (1,...,n-1) this equation can be simplified to:

$$\Lambda = \frac{\prod_{i=1}^{k} \epsilon_i^{s_i}}{\prod_{v=1}^{n} \gamma_{r_v}} = \frac{\prod_{i=1}^{k} \epsilon_i^{s_i}}{\prod_{r=1}^{k-1} \gamma_r^{n_r}} \qquad (3.12)$$

The Conditional Maximum Likelihood estimators (CML) can be derived from this conditional likelihood function. To do this, take logs of both sides, differentiate with respect to all the ϵ_i, and set them equal to zero. This yields

$$\frac{\partial \log \Lambda}{\partial \epsilon_i} = \frac{s_i}{\epsilon_i} - \sum_{r=1}^{k-1} \frac{\gamma_{r-1}^{(i)}}{\gamma_r} \qquad (i=1,\ldots,k) \qquad (3.13)$$

in which the symbol $\gamma_{r-1}^{(i)}$ is used to denote the partial derivative of γ_r with respect to ϵ_i. This derivative is a symmetric function of order r-1 in all parameters except ϵ_i. From (3.13) one arrives at (3.14) which are a set of nonlinear equations in the ϵ_i.

$$s_i = \sum_{r=1}^{k-1} \frac{n_r \epsilon_i \gamma_{r-1}^{(i)}}{\gamma_r} \qquad (i=1,\ldots,k) \qquad (3.14)$$

We can see that since the sum of the s_i equals the sum of the n_r, we must impose some further constraint on the system of equations to allow for a solution. Once again, this can be done in a variety of ways, either by specifying an origin or by specifying a scale.

The great problem in solving the system of nonlinear equations in (3.14) has been the accurate and rapid computation of the symmetric functions and their first and second order partial derivatives. Fischer (1974) presents three useful formulas for the computation of the symmetric functions and their first and second order partial derivatives: (p. 242)

$$\gamma_r = \epsilon_i \gamma_{r-1}^{(i)} + \gamma_r^{(i)} \qquad (3.15)$$

$$r\gamma_r = \sum_{i=1}^{k} \epsilon_i \gamma_{r-1}^{(i)} \qquad (3.16)$$

and (p. 250)

$$\gamma_r(\epsilon_1,\ldots,\epsilon_t) = \gamma_r(\epsilon_1,\ldots,\epsilon_{t-1}) + \epsilon_t \gamma_{r-1}(\epsilon_1,\ldots,\epsilon_{t-1})$$

with $0 \le r \le t$, $t = 1,\ldots,k$. (3.17)

Equations (3.15) and (3.16) can be combined recursively to give a very efficient algorithm for computing the γ_r and the $\gamma_{r-1}^{(i)}$. (Fischer, 1974, pp. 243-244; Gustafsson, 1977,

pp. 30-31). This algorithm is not numerically stable, however, and it usually breaks down when there are more than 20 to 50 items.

Using (3.17) recursively it is, however, possible to devise a numerically stable algorithm for computing the values of the symmetric functions of all orders (Gustafsson, 1977, pp. 31-31), and the derivatives can also be obtained if the algorithm is applied with the parameter value set to zero for the item or items with respect to which the differentiation is made (Fischer, 1974, p. 250). This method allows computations of the symmetric functions for very large sets of items, but it has the drawback that the computations are quite cumbersome and slow when there are many items (more than 50 to 60, say).

However, as was shown by Gustafsson (in press {a}), it is possible to devise an algorithm which is both fast and accurate if (3.17) is used to compute the values of the symmetric functions themselves, and if (3.14) is used to compute the values of the first derivatives. For those items which have extreme parameter values the computations of the first derivatives do break down, but since it is possible to test for numerical accuracy, the derivatives with respect to these items can be recomputed using (3.17) with the parameter value set equal to zero for the item.

Having routines for computing the symmetric functions and their derivatives it is a rather simple matter to solve (3.14), using numerical procedures. One useful method is Newton-Raphson's method (for the details, see Allerup & Sorber, 1977; Andersen, 1972; Fischer, 1974; Wright & Douglas, 1977b). With this method only few iterations are needed, but each iteration requires much computational work since the second derivatives of the symmetric functions must be computed and a (k-1)by(k-1) matrix must be inverted.

Another useful method is based on a simple switching between the right hand side and the left hand side of (3.13) (Fischer, 1974, Gustafsson, 1977; Martin-Löf, 1973; Wright & Douglas, 1977a,b). In this simple iterative method each iteration requires relatively little computational work, but, on the other hand, convergence is slow. However, convergence may be speeded up through using the Aitken extrapolation (Fischer, 1974, p. 245; Fischer & Allerup, 1968; Gustafsson, 1977, p. 35). Usually this extrapolation effects a very considerable saving of iterations, and when it is applied this simple method is in most cases much more effective than Newton-Raphson's method.

It is impossible to give any generally valid guidelines concerning the amount of computer time needed to compute the CML estimates, even for a given number of items, since that is strongly affected by the range of item parameters, and of course also by which particular computer is used. However, when there are no extreme item parameters, relatively little computational work is needed if there is a moderate number of items. For example, on an IBM 360/ 65, the item parameters in a test with 40 items can often be estimated within 3 to 4 seconds of CPU-time, and for a test with 60 items 15 to 10 seconds often suffice.[3]

However, when there are more than 80 to 100 items in the test a large amount of computational work is needed, which is due to the fact that the fastest method of computing the derivatives of the symmetric functions is no longer available; for most of the items the numerical breakdown occurs, which

[3] These estimates were obtained with a FORTRAN IV program, written for the IBM 360/370. A copy of the program written on tape may be obtained at cost from Jan-Eric Gustafsson, Institute of Education, University of Goteborg, Fack, S-431 20 Mölndal, Sweden.

makes it necessary to use the much more cumbersome method based on (3.17).

Estimating Ability

The estimation of person parameters (ability) could be thought of as the dual of the estimation problem just solved for the items. Thus we could set up a system of equations which parallel the item scheme (i.e. determine a conditional likelihood function on item score expressed only in person parameters). This would then yield a set of equations:

$$r_v = \sum_{i=1}^{k} \frac{\theta_v \gamma_{s-1}^{(v)}\{(\theta_v)\}}{\gamma_s\{(\theta_v)\}} \qquad (v=1,\ldots,n) \qquad (3.18)$$

(from Fischer, 1974, p. 240). This system of equations cannot be solved because it is not possible to compute the symmetric functions in the θ_v parameters. If we assume that the usual situation holds, that is, that the number of persons is large in comparison with the number of items, we can treat the estimates of the item parameters as fixed and estimate the person parameters under this assumption. We then get the equations shown in (3.19) to solve:

$$r = \sum_{i=1}^{k} \frac{\theta_r \varepsilon_i}{1+\theta_r \varepsilon_i} \qquad (r=1,\ldots,k-1) \qquad (3.19)$$

This is the same set of equations that were obtained in the unconditional case, except that the subscript v has been changed to r, which is possible since persons having the same raw score are assigned the same ability. These equations are

efficiently solved using Newton-Raphson.

Recently Andersen and Madsen (1977) have presented another approach to make inferences about the person parameters. They define what they call the population likelihood and show that it is possible to estimate the parameters of the distribution of person parameters, assuming a certain distribution function, such as the normal one.

This method has as yet only been used in illustrative examples, but it shows great promise for a wide range of applications. Thus, it may be used to test the hypothesis that the distribution of person parameters is normal, and the estimates of the mean and the variance of the latent distribution would be estimates of the "true mean" and the "true variance." If the mean and the variance of the person parameters estimated from (3.22) are computed for a group of persons these would be estimates of the "observed mean" and the "observed variance." This, of course, makes possible a direct method for estimating the reliability of the test for a certain group of persons, i.e. through dividing the estimate of the "true variance" with the estimate of the "observed variance."

Information Function and Confidence Intervals

The asymptotic standard errors of the CML estimates of the item parameters can be obtained from the inverse of the Hessian, and if Newton-Raphson's method is used to solve (3.14) these are automatically obtained.

The standard errors for the estimates of the parameters can be obtained from the Fisherian Information function. The statistical information in the sample with respect to any parameter Π is defined as:

$$I(\pi) = E\{(\frac{\partial \log \Lambda}{\partial \pi})^2\} \qquad (3.20)$$

where Λ is the likelihood function (see equation 3.12).

Birnbaum (in Lord & Novick, 1968, see Fischer, 1974, p. 294 ff) has shown that the information of item i with respect to the person parameter ξ_v is:

$$I_i(\xi_v) = \frac{\exp(\xi_v - \delta_i)}{\{1+\exp(\xi_v - \delta_i)\}^2} \qquad (3.21)$$

The information of a test $[I_t]$ with respect to the person parameter ξ_v is the sum of the information of each of the k items:

$$I_t(\xi_v) = \sum_{i=1}^{k} \frac{\exp(\xi_v - \delta_i)}{\{1+\exp(\xi_v - \delta_i)\}^2} \qquad (3.22)$$

Similarly the information in the sample with respect to the item parameters $\{I_p[\ \delta_i]\}$ is:

$$I_p(\delta_i) = \sum_{v=1}^{n} \frac{\exp(\xi_v - \delta_i)}{\{1+\exp(\xi_v - \delta_i)\}^2} \qquad (3.23)$$

The maximum likelihood estimates are asymptotically normally distributed with standard error equal to $I^{-\frac{1}{2}}$. Thus, confidence intervals around the item parameters can be

constructed (when the number of examinees is large) in the usual way:

$$\hat{\delta}_i - z_\alpha \sqrt{I_p(\delta_i)^{-1}} \leq \delta_i \leq \hat{\delta}_i + z_\alpha \sqrt{I_p(\delta_i)^{-1}} \qquad (3.24)$$

where Z_α are the critical values obtained from the normal distribution.

When the test is at least of moderate length (more than say, 30 or 40 items) the asymptotic properties should hold sufficiently well for one to make use of them in the determination of confidence intervals around the person parameters. These are:

$$\hat{\xi}_v - z_\alpha \sqrt{I_t(\xi_v)^{-1}} \leq \xi_v \leq \hat{\xi}_v + z_\alpha \sqrt{I_t(\xi_v)^{-1}} \qquad (3.25)$$

Of course these confidence intervals apply only to a randomly chosen person, and not to a particular one (see Lord & Novick, 1968, p. 512).

IV. TESTING GOODNESS OF FIT

The Rasch model is a very strong model with rather stringent assumptions. The desirable consequences of these assumptions are only viable if the assumptions hold. Thus it is crucial to have sensitive tests to determine fit to the model.

On the basis of the UML approach Wright & Panchapakesan (1969) and Mead (1976, Note 2)) have constructed tests of fit.

For these tests the chi-square distribution has been relied upon. The tests have, however, unknown asymptotic properties and simulation studies indicate that even though the means of the distribution match what is expected, the variances may depart substantially (Mead, 1976).

On the basis of the CML approach it is, in contrast, possible to devise tests with known asymptotic properties. There are several such goodness-of-fit tests available for the Rasch model (Andersen, 1973; Martin-Löf, 1973), each of which is sensitive to different threats against the model assumptions. The tests are presented by Gustafsson, (Note 2), and for a fuller treatment of the goodness-of-fit problem than can be afforded here, the reader is directed to this source.

Andersen's Conditional Likelihood Ratio Test

The logarithm of the conditional likelihood function was used earlier and is:

$$\log\Lambda = \sum_{i=1}^{k} s_i \log\varepsilon_i - \sum_{r=1}^{k-1} n_r \log\gamma_r \qquad (4.1)$$

The task of the parameter estimation is to maximize Λ in (4.1). When this has been done, that is when the item parameters for the total sample have been estimated, one inserts them in (4.1) and calculates the maximum of the log likelihood function. This is denoted H_t.

If the model fits, it is expected that the same item parameters should hold in all sub-groups of the person sample. Thus, one estimates the item parameters in all of the k-1 score groups and the values of the log likelihood function. These are then denoted H_r [r = 1,...,k-1] and used

to form the statistic:

$$\log\lambda = H_t - \sum_{r=1}^{k-1} H_r \qquad (4.2)$$

This allows the test of fit by using the property that $-2\log(\lambda)$ is asymptotically chi-square distributed with (k-1)(k-2) degrees of freedom.

Obviously this sort of test has limited application since some 50-100 persons are needed within each score group. Andersen also showed that the k-1 score groups can be pooled into, say, g nonoverlapping groups and the same statistic formed for the g groups. The result is still distributed as chi-square but now with (g-1)(k-1) degrees of freedom.

It is not necessary to divide the sample of persons into groups on the basis of their performance--any disjoint grouping of the sample can be used. But depending on how the grouping is done the test is sensitive to different violations of the model assumptions. If the grouping is made according to level of performance the test is sensitive to variations in the slopes of the ICC's for the items. If another grouping is used, such as according to sex, the test is sensitive to such kinds of multidimensionality which show as item bias, that is, that an item is systematically too easy or difficult for a group of persons.

Martin-Löf's Chi-Square Test

Martin-Löf (1973) has developed a chi-square test for overall goodness of fit based upon the fit within each score group. The logic of his test is as follows:

The number of individuals with raw score r is denoted n_r,

the number in the rth score group who get item i correct is denoted n_{ir}. Thus, the observed proportion of correct answers to item i within score group r is n_{ir}/n_r. The conditional probability that a person with raw score r answers item i correctly is equal to the number of response vectors in which item i is answered correctly divided by the total number of response vectors which have a score of r, that is,

$$P\{A_{vi}=1 \mid r,(\varepsilon_i)\} = \pi_{vi} = \frac{\varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r} \qquad (4.3)$$

Thus, if the model fits, the relation

$$\frac{n_{ir}}{n_r} = \frac{\varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r} \qquad (4.4)$$

should hold for all score groups. Multiplying both sides of this equation by n_r we get an expression for the predicted number of correct responses to each item for each score group. If we define the vector of observed frequencies $q_r' = [n_{1r}, n_{2r}, \ldots, n_{kr}]$, and the corresponding vector of predicted frequencies (from 4.3) t_r, the appropriate test statistic is then:

$$T = \sum_{r=1}^{k-1} \{(q_r)-(t_r)\}'\{((V_r))\}^{-1}\{(q_r)-(t_r)\} \qquad (4.5)$$

in which the matrix V_r is a variance-covariance matrix of

order k-by-k with elements

$$\left.\begin{array}{ll} \dfrac{n_r \varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r} & \text{for } i=j \\ \dfrac{n_r \varepsilon_i \varepsilon_j \gamma_{r-2}^{(i,j)}}{\gamma_r} & \text{for } i\neq j \end{array}\right\} \qquad (4.6)$$

The test statistic T is asymptotically chi-square with (k-1)(k-2) degrees of freedom. If some $n_r = 0$, the summation in (4.5) must be restricted to those score groups that are nonempty (say R of them). If this is done, the degrees of freedom are then (k-1)(R-1). This test is sensitive to variations in the slopes of the ICC's and it is asymptotically equivalent to the Andersen test, when the item parameters in the latter test are estimated within the score groups.

The Martin-Löf Test of Homogeneity of Two Sets of Items

The tests which are sensitive to variations in the slopes of the ICC's may fail to detect multidimensionality such that different groups of items measure different person parameters (Gustafsson, Note 1). However, Martin-Löf (1973) has presented a conditional likelihood ratio test which tests the hypothesis that two groups of items measure the same ability.

To compute the test it is necessary that the items be grouped into two disjoint sets. Let us say that there are k_1 and k_2 items in the two sets, respectively, and that $k_1+k_2 = k$. Furthermore, let $n_{r_1 r_2}$ be the number of persons with raw score r_1 on the first set and raw score r_2 on the second set.

When the item parameters for the total set of k items are estimated, a maximum of the logarithm of the conditional likelihood function is obtained (H_t), and when the item parameters are estimated for each set separately, the corresponding maxima H_1 and H_2 are obtained. The following test statistic can then be formed:

$$\log\lambda = -\sum_{r_1=0}^{k_1}\sum_{r_2=0}^{k_2} n_{r_1 r_2} \log\frac{n_{r_1} n_{r_2}}{n} + \sum_{r=0}^{k} n_r \log\frac{n_r}{n} + H_t - H_1 - H_2 \qquad (4.7)$$

Martin-Löf (1973) has shown that $-2\log(\lambda)$ is approximately chi-square distributed with $k_1 k_2 - 1$ degrees of freedom when n tends toward infinity.

If the items are grouped according to level of difficulty this test is sensitive to variations in levels of person reliability (cf. Lumsden, 1978). But the test can also be applied with the items grouped according to two hypothesized dimensions supposed to be running through the test. In this kind of application, the test of course investigates the hypothesis that the two groups of items measures different abilities.

V. SUMMARY AND DISCUSSION

Of the two estimation procedures discussed above in detail, most researchers have been forced to use the unconditional procedure when applying the Rasch model to tests of more than 20 to 30 items. An algorithm has now been developed, which makes the conditional procedure a feasible alternative to the unconditional method.

A principal advantage of the conditional procedure appears to be the known asymptotic properties of the estimates, which allows the use of the goodness-of-fit tests

described earlier. We therefore recommend that as soon as a thorough analysis of fit of the data to the model is judged important, the conditional procedure, along with these tests, should be used.

Another advantage of the conditional procedure is the availability of the Andersen and Madsen (1977) method for estimating the parameters of the latent population distribution, and for testing hypotheses about this distribution. This methodology will most likely prove very useful in those applications where inferences about groups of persons are intended.

Extensive studies of differences between item difficulties obtained through each method have yet to be done. Most likely, no important practical differences between the methods will be found. The unconditional method is in most cases faster, so when cost is a serious issue there are sometimes strong reasons to prefer this method rather than the conditional one. This can be done profitably in cases when the question of fit is of less importance, either because it can confidently be assumed that the data fit the model, or because the robustness of the model can be relied upon in the solution of practical measurement problems. Furthermore, for very long tests (over 100 items, say) only the unconditional method is feasible.

Developments on the Rasch model have been underway in Europe and the United States for the past two decades. However, mainly due to language problems European work has been little known in the United States. This paper was an attempt to overcome this difficulty.

ACKNOWLEDGMENTS

This research was supported by the Law Enforcement Assistance Administration Grant No. 78-NI-AX-00047, to the Bureau of Social Science Research, Howard Wainer Principal Investigator, and the Board of Regents of the New York State Department of Education in a grant to Touchstone Applied Science Associates, Bertram L. Koslin Principal Investigator. Gustafsson's work has been supported by the Swedish Council for Research in the Humanities and Social Sciences and by the National Board of Education in Sweden. This paper has profited from earlier readings by Bert F. Green, David Thissen and David Weiss, although defects that remain are partially due to the authors not following their advice completely.

REFERENCE NOTES

(1) Gustafsson, J-E. Testing and obtaining fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 8-12, 1979.

(2) Mead, R. Assessing the fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

REFERENCES

Allerup, P. & Sorber, G. The Rasch model for questionaires: with a computer program . Copenhagen: The Danish Institute for Educational Research, 1977.

Andersen, E. B. The solution of a set of conditional estimation equations. Journal of the Royal Statistical Society, 1972, 34, 42-54.

Andersen, E. B. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-140.

Andersen, E. B. Sufficient statistics and latent trait models. Psychometrika, 1977, 42, 69-81.

Andersen, E., & Madsen, M. Estimating the parameters of the latent population distribution. Psychometrika, 1977, 42, 357-374.

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

Fischer, G. H. A further note on estimation in Rasch's measurement model with two categories of answers. Research Bulletin No. 3, Psychological Institute, University of Vienna, 1970.

Fischer, G. H. Einfuehrung in die Theorie psychologischer Tests: Grundlagen und Anwendungen. Bern: Huber, 1974.

Fischer, G. H., & Allerup, P. Rechentechnische Fragen zu Raschs eindimensionalem Modell. In Fischer (ed.) Psychologische Testtheorie. Huber, Bern, 1968.

Fischer, G. H., & Scheiblechner, H. Algorithmen und Programme fuer das probabilistische Testmodell von Rasch. Psychologische Beitraegen, 1970, 12, 23-51.

Gulliksen, H. O. Theory of mental tests. New York: Wiley, 1950.

Gustafsson, J-E. The Rasch model for dichotomous items: theory, applications and a computer program. The Institute of Education University of Goteborg, 1977.

Gustafsson, J-E. A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. Educational and Psychological Measurement, in press, a.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading: Addision-Wesley, 1968.

Lumsden, J. Tests are perfectly reliable. British Journal of Statistical and Mathematical Psychology, 1978, 31, 19-26.

Martin-Löf, P. Statistiska modeller. Anteckningar från seminarier läsåret 1969-70 utarbetade av Rolf Sundberg. 2:a uppl. (Statistical models. Notes from seminars 1969-70 by Rolf Sundberg. 2nd ed.) Institutet för försäkringsmatematik och matematisk statistik vid Stockholms universitet, 1973.

Mead, R. Assessment of fit of data to the Rasch model through analysis of residuals. Unpublished doctoral dissertation, University of Chicago, 1976.

Neyman, J., & Scott, E. L. Consistent estimates based on partially consistent observations. Econometrika, 1948, 16, 1-5.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Cophenhagen: The Danish Institute for Educational Research, 1960.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Scheiblechner, H. A simple algorithm for CML-parameter estimation in Rasch's probabilistic measurement model with two or more categories of answers. Research Bulletin No.5. Psychological Institute, University of Vienna, 1971.

Wright, B. D., & Douglas, G. A. Rasch item analysis by hand. Research Memorandum, No. 21, Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D., & Douglas, G. A. Best procedures for sample-free item analysis. Applied Psychological Measurement, 1977, 1, 281-296. (a)

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

Wright, B. D., & Mead, R. J. BICAL: calibrating items and scales with the Rasch Model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1977.

Wright, B. D., & Panchapakesan, H. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Wright, B. D. and Stone, M. Best Test Design. MESA Press: Chicago, 1979.

END