MAXIMUM LIKELIHOOD ESTIMATION IN GENERALIZED RASCH MODELS

Jan de Leeuw
Department of Data Theory
State University of Leiden
Norman Verhelst
Department of Psychometrics
State University of Utrecht

ABSTRACT

We review various models and techniques that have been proposed for item analysis according to the ideas of Rasch. A general model is proposed that unifies them, and maximum likelihood procedures are discussed for this general model. We show that unconditional maximum likelihood estimation in the functional Rasch model, as proposed by Wright and Haberman, is an important special case. Conditional maximum likelihood estimation as proposed by Rasch and Andersen is another important special case. Both procedures are related to marginal maximum likelihood estimation in the structural Rasch model, which has been studied by Sanathanan, Andersen, Tjur, Thissen, and others. Our theoretical results lead to suggestions for alternative computational algorithms.

KEYWORDS

Latent trait model, Rasch model, incidental parameters, maximum likelihood estimation

INTRODUCTION

We start with a short introduction to general latent trait models. They were introduced, rigorously, by Lawley (1943, 1944). Their use is test theory was stimulated enormously by Lord and Novick (1968), especially perhaps in the chapters written by Birnbaum. In Europe latent trait theory was pioneered by Rasch (1960), and developed most extensively by Fischer (1974) and Andersen (1980). A recent review is Andersen (1983).

We only give the necessary framework. Suppose $\underline{x}_{ij}$ are n x m binary random variables. We use the convention of underlining random variables (Hemelrijk, 1966). We suppose that there exist n additional unobserved or latent random variables $\underline{\xi}_i$, which explain the association between the $\underline{x}_{ij}$. We make a number of assumptions.

A1: The $\underline{\xi}_i$ are independent.

A2: The distribution of $(\underline{x}_{i1},\ldots,\underline{x}_{im})$ only depends on $\underline{\xi}_i$ (and not on other $\underline{\xi}_k$).

A3: The random variables $\underline{x}_{i1},\ldots,\underline{x}_{im}$ are independent given $\underline{\xi}_i$.

A4: The conditional distribution of $\underline{x}_{ij}$, given $\underline{\xi}_i$, is the same for all i.

It follows easily from these assumptions that

$$\text{prob}(\underline{X} = X \mid \underline{\xi} = \xi) = \prod_{i=1}^{n}\prod_{j=1}^{m} \text{prob}(\underline{x}_{ij} = x_{ij} \mid \underline{\xi}_i = \xi_i). \qquad (1)$$

For $\text{prob}(\underline{x}_{ij} = x_{ij} \mid \underline{\xi}_i = \xi_i)$ we can write $\pi_j(\xi_i)$, by assumption A4. Additional specifications are added to define specific models. We need additional specifications of the $\pi_j(\xi)$, which are often called the trace lines or item characteristics, and we need additional assumptions on the distributions of the latent variables $\underline{\xi}_i$.

In this paper we shall investigate models in which the $\underline{\xi}_i$ are distributed on the nonnegative reals. The trace lines are the simple rational functions which characterize the one-parameter logistic or Rasch model. In recent years many different versions of the Rasch model have been proposed, and many different ways to estimate the parameters of the Rasch model have been suggested (Andersen, 1980, Gustafsson, 1980, Kelderman, 1984). We have a Rasch model in which the distribution of each $\underline{\xi}_i$ is a one-point distribution concentrated in ξ_i. For this

model we can estimate the ξ_i, and the item parameters ε_j, by the unconditional or unrestricted maximum likelihood method, which maximizes the likelihood over all n + m parameters. This UML-method has been proposed and studied particularily by Wright and Panchapakesan (1969), Wright and Douglas (1977), and Haberman (1977). Alternatively we can use the fact that in the Rasch model the subject total scores, i.e. $\underline{x}_{i1} + \ldots + \underline{x}_{im}$, are sufficient for the ξ_i. By conditioning on the sufficient statistics we get a conditional likelihood which only depends on the item parameters. Maximizing this conditional likelihood gives conditional maximum likelihood estimates. CML-methods were proposed already by Rasch, but they were studied in considerable detail by Anderson. The book Andersen (1980) has the necessary references, compare also Gustafsson (1980), Wainer, Morgan, and Gustaffson (1980). And finally there are marginal maximum likelihood estimates. Here we assume that the $\underline{\xi}_i$ are identically distributed. Parametric MML estimates are computed by assuming that the distribution of $\underline{\xi}$ belongs to some parametric family. The parameters of the distribution are then estimated jointly with the item parameters. The MML approach was pioneered by Lawley and Bock in the context of probit models, but the first systematic study in the Rasch context is by Andersen and Madsen (1977). Simplified computational methods were proposed by Thissen (1982). Interesting applications are in Sanathanan (1974) and Sanathanan and Blumenthal (1978). Recently several authors have also studied nonparametric marginal maximum likelihood estimation, where nothing is assumed about the distribution of $\underline{\xi}$. We mention Tjur (1982), Cressie and Holland (1983), and Kelderman (1984). Mislevy (1984) is somewhere between parametric and nonparametric. Although the papers of Tjur and of Cressie and Holland are very interesting theoretically, they are not completely satisfactory. These authors discuss an extended random Rasch model, which is seemingly more general than the nonparametric marginal Rasch model, because it does not assume an explicit distribution of $\underline{\xi}$. They discuss estimation in the extended model, without indicating the precise relationship to nonparametric maximum likelihood estimation. We shall show below, that the introduction of the extended model is not necessary.

In our paper we aim to accomplish two things. The various versions of the Rasch model will be unified in a single comprehensive model, of which they are all special cases. And the various estimation procedures will be interpreted as maximum likelihood procedures in this general model. It will turn out that in our framework we can easily bridge the gap

between the extended random model of Tjur and the ordinary marginal or random Rasch model. In fact we can show that the nonparametric marginal maximum likelihoods are identical to the conditional maximum likelihood estimates with probability tending to one.

A GENERAL RASCH MODEL

We now make our discussion of the latent trait model we deal with a bit more precise. The data are an n x m binary matrix X, with $x_{ij} = 1$ if individual i gives the correct answer to item j, and $x_{ij} = 0$ otherwise. The x_{ij} are supposed to be realizations of n x m binary random variables $\underline{x}_{ij}$, whose association is determined by n latent variables $\underline{\xi}_i$. Assumptions A1-A4 of the introduction are used. The latent continuuum is called ability in this context, and $\underline{\xi}_i$ is the ability of individual i. The trace lines are specialized according to the one-parameter logistic or Rasch model, which means that

$$\pi_j(\xi) = \xi\varepsilon_j/(1 + \xi\varepsilon_j). \tag{2}$$

Parameter ε_j (a positive real number) is the easiness of item j. Thus we can also write

$$\text{prob}(\underline{x}_{ij} = x \mid \underline{\xi}_i = \xi) = (\xi\varepsilon_j)^x(1 + \varepsilon\xi_j)^{-1}. \tag{3}$$

If we combine (1) and (3) we find

$$\text{prob}(\underline{x}_i = x \mid \xi\ ) = \{\prod_{j\in I(x)} \varepsilon_j\}\xi^{t(x)} \prod_{j=1}^{m} (1 + \xi\varepsilon_j)^{-1}. \tag{4}$$

In (4) we have written t(x) for the sum of the m elements of the binary vector x, and I(x) is the set of indices with $x_j = 1$. Observe that the conditional specification (4) does not depend on i. Observe also that both ability and difficulty assume only nonnegative values.

We now must relate the conditional core of the model to the observed variables. This is done by assuming that each individual i has its own ability distribution F_i (on the positive half-axis), and that

$$\text{prob}(\underline{x}_i = x) = \int \text{prob}(\underline{x}_i = x \mid \xi)\, dF_i(\xi). \tag{5}$$

This must be combined with (4) to find a complete specification of the distribution of $\underline{x}_i$. First introduce some additional notation. Let

$$\gamma(x,\varepsilon) = \prod_{j\varepsilon I(x)} \varepsilon_j, \tag{6}$$

and

$$\gamma_t(\varepsilon) = \sum_x \{\gamma(x,\varepsilon) \mid t(x) = t\}. \tag{7}$$

Also, if $t(x) = t$,

$$\pi_\varepsilon(x|t) = \gamma(x,\varepsilon)/\gamma_t(\varepsilon). \tag{8}$$

If $t(x) \neq t$, then $\pi_\varepsilon(x|t) = 0$. The reason for introducing this notation is as follows. Suppose $\underline{t}_i$ is the sum of the elements of $\underline{x}_i$. Then (4) shows immediately that

$$\text{prob}(\underline{x}_i = x \mid \underline{t}_i = t) = \pi_\varepsilon(x|t), \tag{9}$$

which does not depend on the F_i but only on the difficulties of the items. We also define

$$\pi_\varepsilon(t|\xi) = \gamma_t(\varepsilon)\xi^t \prod_{j=1}^{m} (1 + \xi\varepsilon_j)^{-1}, \tag{10}$$

which is motivated by

$$\text{prob}(\underline{t}_i = t \mid \xi) = \pi_\varepsilon(t|\xi) \tag{11}$$

and

$$\pi_{i,\varepsilon}(t) = \int \pi_\varepsilon(t|\xi)\, dF_i(\xi), \tag{12}$$

which is, of course, the distribution of the $\underline{t}_i$. Notation (6)-(12) makes life really simple. We can write (4) as

$$\text{prob}(\underline{x}_i = x \mid \xi) = \pi_\varepsilon(x|t)\pi_\varepsilon(t|\xi), \tag{13}$$

and (5) as

$$\text{prob}(\underline{x}_i = x) = \pi_\varepsilon(x|t)\pi_{i,\varepsilon}(t). \tag{14}$$

In both (13) and (14) we suppose, of course, that $t = t(x)$. The final assumption we make, to complete our model, is that all vectors $\underline{x}_i$ are independent. From (14)

$$\text{prob}(\underline{X} = X) = \prod_{i=1}^{n} \pi_\varepsilon(x_i|t(x_i)) \prod_{i=1}^{n} \pi_{i,\varepsilon}(t(x_i)). \tag{15}$$

SPECIAL CASES

The original model, proposed by Rasch (1960), and studied most completely by many later authors, is the special case of our general model in which the F_i are step-functions with a single step. They step from zero to one at the point ξ_i. Thus, from (12),

$$\pi_{i,\varepsilon}(t) = \pi_\varepsilon(t|\xi_i). \tag{16}$$

This model could be called the fixed-score model, or the functional model, using analogies with factor analysis and linear errors-in-variables models.

There is also a random-score or structural version of our model, in which we merely suppose that $F_i = F$. Thus individuals are all sampled from the same distribution. We can now write $\pi_\varepsilon(t)$ for $\pi_{i,\varepsilon}(t)$, otherwise there are no simplifications. The structural model has been proposed in the probit context by Lawley in the forties, but for the Rasch model the first systematic study was by Andersen and Madsen (1977), and by Sanathanan (1974), Sanathanan and Blumenthal (1978). Recently the random-effects model has been rapidly gaining in popularity (Tjur, 1982, Cressie and Holland, 1983, Kelderman, 1984).

These two classical cases are by no means the only ones which are interesting. Analysis of the structural model, for instance, can be subdivided into the case in which F is known, the case in which F is known to belong to a parametric family, and the case in which F is completely unknown. We can distinguish similar special cases if the F_i are not assumed to be equal. A nice example is suggested briefly below.

Suppose the F_i are logistic with common variance. Thus

$$F_i(\xi) = \xi\theta_i/(1 + \xi\theta_i). \tag{17}$$

It is possible now to compute $\pi_{i,\varepsilon}(t)$ in closed form. Suppose all ε_j are different. We first use the general partial fraction formula

$$\pi_{i,\varepsilon}(t) = \gamma_t(\varepsilon) \sum_{j=1}^{m} \kappa_j^t(\varepsilon) \int (1 + \xi\varepsilon_j)^{-1} \, dF_i(\xi), \tag{18}$$

with

$$\kappa_j^t(\varepsilon) = (-1)^t \, \varepsilon_j^{m-(t+1)} \prod_{\ell \neq j} (\varepsilon_j - \varepsilon_\ell)^{-1}. \tag{19}$$

Moreover, if $\theta_i \neq \varepsilon_j$,

$$\int (1 + \xi\varepsilon_j)^{-1} \, dF_i(\xi) = \theta_i/(\theta_i - \varepsilon_j) - \theta_i\varepsilon_j(\ln \theta_i - \ln \varepsilon_j)/(\theta_i - \varepsilon_j)^2. \quad (20)$$

Combining (18)(19)(20) gives the required result. It is not very simple, but it does give an explicit expression for the distribution of $\underline{t}_i$ in a fairly interesting case. This generalizes a result in De Leeuw (1973, page 50-52). Other special cases can be developed by assuming other forms of F_i, but for these other special cases either F_i is not very realistic or no closed form solution results. It seems that (17) is a rather natural choice.

Although the parametric assumptions on the F_i lead to a great deal of useful data reduction, we shall concentrate in this paper on nonparametric versions of the Rasch model in which nothing is specified about the F_i. Except, of course, that they must be distribution functions on the nonnegative real numbers. The non-parametric models seem somewhat less arbitrary to us, because they make fewer specific choices. On the other hand they use a very large, in principle infinite, number of unknown parameters, and it is possible that as a consequence of this their statistical properties detoriate. In the next sections we shall show that in the Rasch model efficient estimation of the structural parameters is still possible, even in models with completely general F_i. The fact that this sort of estimation can conveniently be carried out is to a large extent specific for the Rasch model, because it depends on existence of a factorization which causes what Rasch calls <u>specific objectivity</u> (Rasch, 1961, 1966, 1977). Although this factorization is certainly convenient, its importance has been exaggerated greatly by some authors. We also abandon the perfect symmetry of the classical Rasch model, because we parametrize items by real parameters and individuals by distribution functions. This is, of course, just another 'variation on a theme by Thurstone' (Lumsden, 1980).

THE LIKELIHOOD FUNCTION

We are interested in maximum likelihood estimation in this paper. Thus we now give a convenient expression for the loglikelihood function of our general Rasch model. By taking logarithms in (15) we see that the loglikelihood can be decomposed in two parts. Thus

$$L(\varepsilon,F) = \sum_{i=1}^{n} \ln \pi_{\varepsilon}(x_i|t(x_i)) + \sum_{i=1}^{n} \ln \pi_{i,\varepsilon}(t(x_i)), \tag{21}$$

which we write as

$$L_T(\varepsilon,F) = L_C(\varepsilon) + L_P(\varepsilon,F). \tag{22}$$

The <u>total</u> loglikelihood is the sum of the <u>conditional</u> loglikelihood, which only depends on ε, and the <u>population</u> loglikelihood, which depends on both ε and F. We now derive some simplifications. The first one is

$$L_C(\varepsilon) = \sum_{t=0}^{m} n_t \sum_{x} \{p_{x|t} \ln \pi_{\varepsilon}(x|t) \mid t(x)=t \}, \tag{23}$$

which n_t the observed number of individuals with total score t, and with $p_{x|t}$ the proprotion of these individuals who have profile x, where $t(x) = t$. This shows that the conditional loglikelihood is like a product multinomial loglikelihood, with one factor for each value of t. Of course for $t = 0$ and $t = m$ there is no contribution to (18), and thus we can also sum for $t = 1$ to $t = m - 1$. A similar simplication is possible for the marginal likelihood. We find

$$L_P(\varepsilon,F) = \sum_{t=0}^{m} n_t \ln \pi_{F,\varepsilon}(t), \tag{24}$$

with

$$\pi_{F,\varepsilon}(t) = n_t^{-1} \sum \{\pi_{i,\varepsilon}(t) \mid t(x_i) = t\}. \tag{25}$$

In other words

$$\pi_{F,\varepsilon}(t) = \int \pi_{\varepsilon}(t|\xi)\, dF_t(\xi), \tag{26}$$

with

$$F_t = n_t^{-1} \sum \{F_i \mid t(x_i) = t\}. \tag{27}$$

It follows directly form this representation that we can never hope to estimate the individual F_i, only their averages F_t.

The simplifications for the structural and functional Rasch models in which the F_i are equal or one-point distributions are obvious.

UNRESTRICTED MAXIMUM LIKELIHOOD ESTIMATION

If we maximize $L_T(\varepsilon,F)$ over the F_i, it follows directly from (12) that the optimum F_i are one-point distributions. This also follows for the optimal F_t from (26). Thus we may as well assume in the unrestricted case that we are in the functional model in which each individual is characterized by his ability ξ_i, which is the value where F_i jumps from zero to one. Unrestricted maximum likelihood estimation in the general model amounts to the same thing as unrestricted maximum likelihood model in the classical functional Rasch model.

These unrestricted (also called unconditional) maximum likelihood estimates in the Rasch model have well-known problems. Andersøn (1973) has shown that they are not consistent in the asymptotic case with $n \to \infty$ and m fixed. Haberman (1977), following a suggestion of Lord (1975), has shown that they are consistent if both $n \to \infty$ and $m \to \infty$, provided $m^{-1} \ln n \to 0$. Because of these complications Andersøn (1970), following suggestions of Rasch, has suggested the conditional maximum likelihood estimates of ε, which maximize $L_C(\varepsilon)$. They can be thought of either as estimates in a conditional likelihood model, where we condition on the total scores of the individuals, or as approximate total likelihood estimates. In this last interpretation (Andersøn and Madsen, 1977) we complete the estimation process by maximizing $L_P(\hat{\varepsilon},F)$ over F, with $\hat{\varepsilon}$ the conditional maximum likelihood estimates of ε. Again the optimum F is one-step, where $\hat{\xi}_t$ solves the equation

$$\sum_{j=1}^{m} \hat{\xi}\hat{\varepsilon}_j/(1 + \hat{\xi}\hat{\varepsilon}_j) = t. \tag{28}$$

The conditional maximum likelihood estimates are consistent, in the model with $n \to \infty$ and m fixed, but the $\hat{\xi}_t$ (or $\hat{F}_t$ or even $\hat{F}_i$) do not have clear cut statistical properties. They are, of course, asymptotically normal and satisfy the population version of (28), but this is not at all like consistency. The functional model, and even more so the general model, are hampered by the problem of incidental parameters (Neyman and Scott, 1948, Kiefer and Wolfowitz, 1956). They cause inconsistencies and other forms of trouble. The conditional maximum

likelihood estimates avoid the problem of inconsistency, but they involve the somewhat artificial operation of conditioning, and they are computationally rather demanding. It turns out to be possible to justify conditional maximum likelihood estimates in a somewhat different way. This derivation also indicates some possible computational applications.

ESTIMATION IN THE STRUCTURAL MODEL

In this section we prove that the maximum likelihood estimates of ε in the structural model, in which $F_i = F$ for all i, are identical to the conditional maximum likelihood estimates discussed in the previous section. This result seems to be new. In recent papers dealing with the structural model Tjur (1982) and Cressie and Holland (1983) discuss the extended random model, in which the likelihood function $L_C(\varepsilon) + L_P(\pi)$ is maximized, where

$$L_P(\pi) = \sum_{t=0}^{m} n_t \ln \pi_t, \tag{29}$$

and the π_t are unrestricted (except for the fact that they must be nonnegative numbers adding up to unity). The extended model is seen as being quite different from the structural Rasch model, however. Tjur remarks that the $\pi_{F,\varepsilon}(t)$ 'are complicated functions of the unknown parameters $\varepsilon_1,\ldots,\varepsilon_m$ and F, and an attempt to maximize the likelihood directly as a function of $(\varepsilon_1,\ldots,\varepsilon_m,F)$ would hardly be successful.' (1982, page 24). Cressie and Holland remark that the estimates of $(\varepsilon_1,\ldots,\varepsilon_m,\pi_0,\ldots,\pi_m)$ in the extended model are consistent and asymptotically normal, but 'possibly somewhat inefficient' for the structural Rasch model (1983, page 137). We show that the marginal maximum likelihood estimates of ε are identical (with probability tending to one) to the conditional maximum likelihood estimates, which implies directly that they have the same asymptotic normal distribution. Thus conditional maximum likelihood estimates are efficient in the structural model. We also show that F cannot be estimated consistently, unless we specify it in more detail.

The key result with which we start is, that that maximum of $L_T(\varepsilon,F)$ is always less than or equal to the maximum of $L_C(\varepsilon)$ plus the maximum of $L_P(\pi)$. It is equal if and only if we can find F in such

a way that $L_p(\hat{\varepsilon},F) = p_t$, with $p_t = n_t/n$, and with $\hat{\varepsilon}$ the conditional maximum likelihood estimates. If we can find F such that these equations are satisfied, then $\hat{\varepsilon}$ are marginal maximum likelihood estimates, and F is also maximum likelihood. Let us analyse these equations a bit more in detail. They are

$$\gamma_t(\hat{\varepsilon}) \int \xi^t \prod_{j=1}^{m} (1 + \xi\hat{\varepsilon}_j)^{-1} \, dF(\xi) = p_t. \tag{30}$$

When is (30) solvable for F ? The functions that are integrated in (30) are of the form $\alpha(t)\beta(\xi)\xi^t$. Thus they are a simple rescaling of the monomials ξ^t, and consequently they form a Tchebycheff system (Karlin and Studden , 1966, chapter I, Krein and Nudel'man, 1977, chapter II). Thus (30) defines a generalized moment problem (or Tchebycheff moment problem) on the nonnegative real line. Such problems are analyzed in detail in Karlin and Studden (1966, chapter V) and Krein and Nudel'man (1977, chapter V). We follow Krein and Nudel'man, and first reduce (30) in a simple way to a power moment problem (or Stieltjes moment problem). In the first place (30) is solvable if and only if

$$\int \xi^t \prod_{j=1}^{m} (1 + \xi\hat{\varepsilon}_j)^{-1} \, dF(\xi) = p_t/\gamma_t(\hat{\varepsilon}) \tag{31}$$

is solvable. Now consider the functions

$$u_t(\xi) = \xi^t \prod_{j=1}^{m} (1 + \xi\hat{\varepsilon}_j)^{-1}. \tag{32}$$

They satisfy $u_t(\xi) \geq 0$, and all $u_t(\xi)$ tend to a finite limit if $\xi \to \infty$. Thus condition 1.1 of Krein and Nudel'man (1977, page 173) is satisfied, and (31) is solvable if and only if the sequence $p_t/\gamma_t(\hat{\varepsilon})$ is positive. This means that we must have $\sum a_t p_t/\gamma_t(\hat{\varepsilon}) \geq 0$ for all a_t such that $\sum a_t u_t(\xi)$ is a positive function on the half line. But clearly $\sum a_t u_t(\xi)$ is positive if and only if $\sum a_t \xi^t$ is positive. Thus (31) is solvable if and only if there exists a nondecreasing G such that

$$\int \xi^t \, dG(\xi) = p_t/\gamma_t(\hat{\varepsilon}) \tag{33}$$

is solvable, and this is, of course, a power moment problem.

Necessary and sufficient conditions for the solvability of the power moment problem are well known (Karlin and Studden, 1966, chapter V, section 10, or Krein and Nudel'man, 1977, chapter V, page 175-176). These results have been used earlier in the context of (extended) Rasch models by Cressie and Holland (1983) and Kelderman (1984). We briefly recapilute them here. Define two matrices $A(\hat{\varepsilon})$ and $B(\hat{\varepsilon})$, both symmetric. If m is even, then $A(\hat{\varepsilon})$ is of order $\frac{1}{2}m + 1$ and $B(\hat{\varepsilon})$ is of order $\frac{1}{2}m$. We have $a_{st}(\hat{\varepsilon}) = p_{s+t}/\gamma_{s+t}(\hat{\varepsilon})$ for $s,t=0,\ldots,\frac{1}{2}m$ and $b_{st}(\hat{\varepsilon}) = p_{s+t+1}/\gamma_{s+t+1}(\hat{\varepsilon})$ for $s,t=0,\ldots,\frac{1}{2}m-1$. If m is odd, then $A(\hat{\varepsilon})$ and $B(\varepsilon)$ both have order $\frac{1}{2}(m + 1)$. Now $a_{st}(\hat{\varepsilon}) = p_{s+t}/\gamma_{s+t}(\hat{\varepsilon})$ for $s,t=0,\ldots,\frac{1}{2}(m-1)$ and $b_{st}(\hat{\varepsilon}) = p_{s+t+1}/\gamma_{s+t+1}(\varepsilon)$ for $s,t=0,\ldots,\frac{1}{2}(m-1)$. The power moment problem (33) is solvable for G if and only if $A(\hat{\varepsilon})$ and $B(\hat{\varepsilon})$ are both positive semi-definite. If they are both positive definite, then a solution without mass at infinity exists. If a solutions exists with infinitely many points of increase, and without mass at infinity, then $A(\hat{\varepsilon})$ and $B(\hat{\varepsilon})$ are positive definite.

It follows directly from our discussion above that the nonparametric marginal maximum likelihood estimates of the ε_j are equal to the conditional maximum likelihood estimates $\hat{\varepsilon}_j$ if the two matrices $A(\hat{\varepsilon})$ and $B(\hat{\varepsilon})$ are positive semidefinite. But we can go further than this. Suppose the structural Rasch model, with F a proper distribution function with an infinite number of points of increase, is true. Then $A(\varepsilon)$ and $B(\varepsilon)$, the population values of the matrices, are positive definite. Because both p and $\hat{\varepsilon}$ are consistent estimates of their population values, it follows that $A(\hat{\varepsilon})$ and $B(\hat{\varepsilon})$ converge in probability to $A(\varepsilon)$ and $B(\varepsilon)$ if $n \to \infty$. Thus the probability that they are positive definite tends to one if n tends to infinity. If the Rasch model holds, then nonparametric marginal maximum likelihood estimates are equal to conditional maximum likelihood estimates with probability tending to one. This implies, of course, that they have the same aymptotic normal distribution, and they are both efficient in the structural Rasch model. This constitues our main result, because it shows that the extended Rasch model of Tjur and Cressie & Holland is asymptotically identical to the structural Rasch model.

The situation is far less satisfactory with respect to the estimation of F. In fact F cannot be determined uniquely by maximum likelihood, unless we make additional specifications. If we assume too much, for instance a parametric family, then marginal maximum likelihood may

become quite different from conditional maximum likelihood, as the example with a logistic shift for each individual outlined in an earlier section shows. It is possible, however, to specify that F must be canonical. The theory of canonical representations for moment problems on the positive half-line is developed by Karlin and Studden (1966, chapter V, section 4), for the Stieltjes power moment problem even more details are given by Krein and Nudel'man (1977, chapter V, section 4). The principal solutions are step functions. If m is odd they have $(m + 1)/2$ steps at different points, if m is even they have $(m + 2)/2$ steps at different points, with the first point equal to zero. This follows from the discussion in Karlin and Studden, if we verify their condition that the functions u_t in (32) form a type II system. For this it suffices to see that $u_t(\xi)/u_m(\xi)$ with $t < m$ converges to zero if ξ converges to infinity. We are only interested in the lower principal representation, because the upper one places mass at infinity. This lower representation is unique, both with respect to location and size of the steps. Thus if we are prepared to assume that F is a step function, with the required number of steps, then F can be estimated consistently. But there is no reason to believe that F has this highly artificial form in real live situations. It would perhaps be nicer to estimate the set of all solutions. By adapting the notion of consistency to set-valued functions one can prove that this set is estimated consistently. Interested readers are strongly advised to consult the books of Karlin and Studden, and Krein and Nudel'man, and the literature they cite. There is a veritable goldmine of details in these works, of which some are certainly relevant for the further study of the Rasch model.

We recapitulate the major result of this section briefly. If we assume that the structural Rasch model (with all F_i equal to F, and with F a nondegenerate proper distribution function) is true, then the marginal maximum likelihood estimates of the items parameters are equal to the conditional maximum likelihood estimators with probability one. Thus the CML estimates are efficient in the structural model. This does not imply, however, that the MML estimates are efficient in the functional model, in fact is does not even imply that the MML estimates are consistent in the functional model. The functional model is inherently more complicated than the structural model, and it is already non-standard to define consistency and efficiency in

this model. This asymmetry must be kept in mind. CML estimates behave nicely both in the functional and in the structural model. In fact they even behave nicely in the general model with different F_i. MML estimates only behave nicely if the structural model is true.

ADDITIONAL THEORY

Up to now we have concentrated on the maximum likelihood estimation in the functional and structural Rasch models. We have shows that they can both be interpreted as marginal maximum likelihood methods in particular versions of the general Rasch model, and we have investigated their relationship with the conditional maximum likelihood method. Our analysis also has some theoretical applications, that are interesting are deserve some emphasis.

The first result is not new, because it has already been derived by Cressie and Holland (1983). We repeat it, because it seems important, because it follows directly from our computations, and because our proof is slightly different from the proof in Cressie and Holland. The result is, that our computations illustrate theorem 2 of Cressie and Holland. They give necessary and sufficient conditions for a set of observed profile probabilities to satisfy the structural Rasch model for some F. The conditions are that the conditional distributions on the scores groups can be fitted by the log-linear model of Tjur, Cressie and Holland, and Kelderman. This defines a system of linear equations on the log-conditional-probabilities. The second condition is that the solution of the linear system produces matrices A and B that are positive semi-definite. If they are positive-definite then a proper solution (with no mass at infinity) exists.

Our second theoretical result again makes Cressie and Holland more precise. They pay quite a lot of attention in their section 2 to the degrees of freedom problem in fitting the structural model. It seems, of course, as if the structural model has an infinite number of parameters, because F is not specified. The theory of principal representations shows us what the exact dimensionality is. The number of free parameters turns out to be exactly $2m - 1$, and thus chi-square testing the fit of the structural model has $2^m - 2m$ degrees of freedom. The same conclusion was reached by Cressie and Holland, who started from the extended model and observed that the positivity constraints on A and B do not reduce dimensionality. But, using the lower principal representation, we

can compute the dimensionality directly. If m is odd there are (m + 1)/2 points of increase (knots), there are (m + 1)/2 masses (jumps), and there are m item parameters. The masses add up to one, and one knot can be set equal to one for normalization. Thus there are $\frac{1}{2}(m - 1) + \frac{1}{2}(m - 1) + m = 2m - 1$ free parameters. If m is even there are (m + 2)/2 knots and (m + 2)/2 masses. The smallest knot is zero, one of the others can be set equal to one. The masses add up to one again. Thus the number of free parameters is $(\frac{1}{2}m - 1) + \frac{1}{2}m + m = 2m - 1$.

COMPUTATIONAL APPLICATIONS

Computing CML estimates is generally not an easy task. Compare Wainer, Morgan, and Gustaffson (1980), Gustaffson (1979, 1980) for a review of the currently best implementations. Some possibilities for improvement have recently been suggested by Jansen (1984) and by Verhelst, Glas, and Van der Sluis (1984), but the computations remain rather complicated. UML estimates are very easy to compute, but we have seen that they have some undesirable properties from a theoretical point of view. The results of the previous sections make it possible to compute CML estimates, or at least estimates which are asymptotically equivalent to them, by solving a finite mixture problem (Everitt and Hand, 1981, Redner and Walker, 1984). Of course we can only apply the result if we assume that the structural Rasch model is true, in case of the functional model our results do not apply.

From our previous considerations it follows that we must maximize the likelihood function

$$L_T(\varepsilon,\xi,\theta) = \sum_{j=1}^{m} s_j \ln \varepsilon_j + \sum_{t=0}^{m} n_t \ln \sum_{\nu} \theta_\nu \xi_\nu^t \prod_{j=1}^{m} (1 + \xi_\nu \varepsilon_j)^{-1}, \qquad (34)$$

with s_j the total number of correct responses for item j, and with θ_ν the sizes of the steps of the canonical F (i.e. positive numbers adding up to one). This is a simple function of 2m - 1 free parameters, which can be maximized quite simply in various ways. We shall not go into details here, because this paper is not concerned with computation as such. By using the EM-algorithm, as in other marginal maximum likelihood approaches (Bock and Aitkin, 1981, Thissen, 1982, Mislevy, 1984), we can reduce the problem to a sequence of unconstrained maximum likelihood problems, i.e. to iterative proportional fitting. This is very convenient on small computers, but convergence could very well be slow

(Redner and Walker, 1984). We have the familiar trade-off here between ease of computation and speed of convergence. If we want rapid convergence we can use Newton-Raphson iterations, which are also quite simple for (34). A combination of the two methods which starts with EM and ends with Newton-Raphson may prove to be the best general procedure. All this must be investigated more in detail, however, and compared with existing conditional maximum likelihood methods.

We have concentrated on nonparametric marginal maximum likelihood estimation in the later parts of this paper, because it seemed theoretically more interesting. We must not forget, however, that parametric assumptions on F are currently more popular. They make it possible to estimate F consistently in general, but of course they require a fairly precise specification of F. The model derived in (17)-(20) seems worthy of further investigation, because it seems to be rather simple. It also only involves 2m parameters, and can be used as an interesting alternative to the structural model.

REFERENCES

Andersen, E.B. (1973) Conditional Inference and Models for Measuring. Copenhagen, Mental Hygiejnisk Forlag.

Andersen, E.B. (1980) Discrete Statistical Models with Social Science Applications, Amsterdam, North Holland Publishing Co.

Andersen, E.B. (1983) Latent trait models, Journal of Econometrics, 22, 215-228.

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

Bock, R.D. & Aitkin, M.A. (1981) Marginal maximum likelihood estimation of items parameters: application of an EM-algorithm, Psychometrika, 46, 443-460.

Cressie,N. & Holland, P.W. (1983) Characterizing the manifest probabilities of latent trait models, Psychometrika, 48, 129-141.

De Leeuw, J. (1973) Canonical Analysis of Categorical Data, Leiden, Psychological Institute. Reissued 1984, Leiden, DSWO-press.

Everitt, B.S. & Hand, D.J. (1981) Finite Mixture Distributions, London, Chapman and Hall.

Gustaffson, J.E. (1979) PML: a Computer Program for Conditonal Estimation and Testing in the Rasch Model for Dichotomous Items, Göteborg, Reports from the Institute of Education no 85.

Gustaffson, J.E. (1980) Testing and obtaining fit of data to the Rasch model, British Journal of Mathematical and Statistical Psychology, 33, 205-233.

Haberman, S.J. (1977) Maximum likelihood estimates in exponential response models, Annals of Statistics, 5, 815-841.

Hemelrijk, J. (1966) Underlining random variables, Statistica Neerlandica, 20, 1-8.

Jansen, P.G.W. (1984) Computing the second-order derivatives of the symmetric functions in the Rasch model, Kwantitatieve Methoden, 13, 131-147.

Karlin, S. & Studden, W.J. (1966) Tchebysheff Systems: with Applications to Analysis and Statistics, New York, Wiley.

Kelderman, H. (1984) Loglinear Rasch model tests, Psychometrika, 49, 223-245.

Kiefer, J. & Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Annals of Mathematical Statistics, 27, 887-906.

Krein, M.G. & Nudel'man, A.A. (1977) The Markov Moment Problem and Extremal Problems, Providence, American Mathematical Society.

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

Lawley, D.N. (1944) The factorial analysis of multiple item tests, Proceedings of the Royal Society of Edinburgh, 62, 74-82.

Lord, F.M. (1975) Consistent Estimation when Number of Variables and Number of Parameters Increases without Limit, Princeton, Educational Testing Service.

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

Lumsden, J. (1980) Variations on a theme by Thurstone, Applied Psychological Measurement, 4, 1-7.

Mislevy, R.J. (1984) Estimating latent distributions, Psychometrika, 49, 359-381.

Neyman, J. & Scott, E.L. (1948) Consistent estimates based on partially consistent observations, Econometrica, 16, 1-32.

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

Rasch, G. (1961) On general laws and the meaning of measurement in psychology, Proceedings Fourth Berkeley Symposium in Mathematical Statistics and Probability, volume 5, 321-333.

Rasch, G. (1966) An informal report on a theory of objectivity in comparisons, Proceedings NUFFIC 1966 Summer Session, Den Haag.

Rasch, G. (1977) On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy, 14, 58-94.

Redner, R.A. & Walker, H.F. (1984) Mixture densities, maximum likelihood and the EM-algorithm, SIAM review, 26, 195-239.

Sanathanan, L. (1974) Some properties of the logistic model for dichotomous response, Journal of the American Statistical Association, 69, 744-749.

Sanathanan, L. & Blumanthal, S. (1978) The logistic model and estimation of latent structure, Journal of the American Statistical Association, 73, 794-799.

Thissen, D. (1982) Marginal maximum likelihood estimation for the one-parameter logistic model, Psychometrika, 47, 175-186.

Tjur, T. (1982) A connection between Rasch's item analysis model and a multiplicative Poisson model, Scandinavian Journal of Statistics, 9, 23-30.

Verhelst, N.D., Glas, C.A.W., & Van der Sluis, A. (1984) Estimation problems in the Rasch model: the basic symmetric functions. Computational Statistics Quarterly, in press.

Wainer, H., Morgan, A., & Gustaffson, J.E. (1980) A review of estimation procedures for the Rasch model with an eye toward longish tests, Journal of Educational Statistics, 5, 35-64.

Wright, B.D. & Douglas, G.A. (1977) Best procedures for sample-free item analysis, Applied Psychological Measurement, 1, 281-295.

Wright, B.D. & Panchapakesan, N. (1969) A procedure for sample-free item analysis, Educational and psychological Measurement, 29, 23-48.