

*Journal of Educational Statistics*
*Fall, 1986, Volume 11, Number 3, pp. 183–196*

# MAXIMUM LIKELIHOOD ESTIMATION IN GENERALIZED RASCH MODELS

JAN DE LEEUW
State University of Leiden
and
NORMAN VERHELST
National Institute for Educational Measurement (Cito), Arnhem

KEY WORDS. *Latent trait model, Rasch model, incidental parameters, maximum likelihood estimation.*

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.

We start with a short introduction to general latent trait models. They were introduced rigorously by Lawley (1943, 1944). Their use in test theory was stimulated enormously by Lord and Novick (1968), especially 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 $n \times m$ 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

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

For $\text{prob}(\underline{x}_{ij} = 1|\underline{\xi}_i = \xi_i)$ we can write $\pi_j(\xi_i)$ by assumption A4. Other 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 characteristic curves*, 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 that 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 $\xi_i$. For this model we can estimate the $\xi_i$ and the item parameter $\varepsilon_j$ by the unconditional or *unrestricted maximum likelihood (UML) method*, which maximizes the likelihood over all $n + m$ parameters. The UML method has been proposed and studied particularly 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, that is, $\underline{x}_{i1} + \cdots + \underline{x}_{im}$, are sufficient for the $\xi_i$. By conditioning on the sufficient statistics we get a conditional likelihood that only depends on the item parameters. Maximizing this conditional likelihood gives *conditional maximum likelihood (CML) estimates.* CML methods were proposed already by Rasch, but they were studied in considerable detail by Andersen (1973). (Andersen, 1980, has the necessary references; cf. also Gustafsson, 1980; Wainer, Morgan, & Gustafsson, 1980.)

Finally, there are *marginal maximum likelihood (MML) 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 (Andersen & Madsen, 1977; Thissen, 1982). (See also Sanathanan, 1974, and Sanathanan and Blumenthal, 1978, for some interesting applications.)

Recently several authors have also studied nonparametric MML estimation, where nothing is assumed about the distribution of $\underline{\xi}$. We mention Tjur (1982), Cressie and Holland (1983), and Kelderman (1984). In the approach of Tjur and Cressie and Holland, the item parameters and the first $m$ moments of the distribution of $\underline{\xi}$ are estimated by a maximum likelihood procedure. In Tjur's

approach, these moments are not recognized as such, so no inequality constraints are put on them. Cressie and Holland give the necessary constraints to be put on the estimated moments for a distribution function of the latent variable to exist. The model discussed by Tjur is denoted the extended random Rasch model.

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 the versions are all special cases. And the various estimation procedures will be interpreted as MML 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 MML estimates are identical to the CML estimates with probability tending to one.

## A General Rasch Model

We now make more precise our discussion of the latent trait model we deal with. The data are an $n \times m$ 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 $n \times m$ 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 continuum 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 $\varepsilon_j$ (a positive real number) is the *easiness* of item $j$. Thus we can also write

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

If we combine Equations 1 and 3 we find

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

In Equation 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 in Equation 4 does not depend on $i$. Observe also that both ability and easiness 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

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

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

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

and

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

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 Equation 4 shows immediately that

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

which does not depend on the $F_i$ but only on the easiness 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

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

and

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

which is, of course, the distribution of $\underline{t}_i$. Notation in Equations 6–12 makes life really simple. We can write Equation 4 as

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

and Equation 5 as

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

The final assumption we make, to complete our model, is that all vectors $\underline{x}_i$ are independent. From Equation 14

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

### *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 $\xi_i$. Thus, from Equation 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 was proposed in the probit context by Lawley in the 1940s, but for the Rasch model the first systematic studies were by Andersen and Madsen (1977), Sanathanan (1974), and Sanathanan and Blumenthal (1978). Recently the random-score model has been rapidly gaining in popularity (Cressie & Holland, 1983; Kelderman, 1984; Tjur, 1982).

These two classical cases are by no means the only ones that 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 given by the case where the $F_i$ are logistic with common variance, where it is possible to compute $\pi_{i,\varepsilon}(t)$ in closed form. A detailed discussion of this model will be reported elsewhere.

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 nonparametric 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 perhaps as a consequence their statistical properties deteriorate.

In the following 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 the existence of a factorization that causes what Rasch (1961, 1966, 1977) calls *specific objectivity*.

Although this factorization is certainly convenient, its importance has been greatly exaggerated 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, of course, is just another "variation on a theme by Thurstone" (Lumsden, 1980).

## The Likelihood Function

We are interested in maximum likelihood estimation. Thus we now give a convenient expression for the loglikelihood function of our general Rasch model. In a first step, we follow Andersen and Madsen (1977) and decompose the loglikelihood in two parts. By taking logarithms of Equation 15 we see that

$$L(\varepsilon,F) = \sum_{i}^{n} \ln \pi_{\varepsilon}(x_i|t_i) + \sum_{i}^{n} \ln \pi_{i\varepsilon}(t_i), \tag{17}$$

which we can write as

$$L_T(\varepsilon,F) = L_C(\varepsilon) + L_P(\varepsilon,F).$$

The *total* loglikelihood is the sum of the conditional loglikelihood, which only depends on $\varepsilon$, and the population loglikelihood, which depends on both $\varepsilon$ and $F$. We now consider two simplifications. The first one is

$$L_C(\varepsilon) = \sum_{t=0}^{m} n_t \sum_{\substack{x \\ t(x)=t}} p_{x|t} \ln \pi_{\varepsilon}(x|t), \tag{18}$$

with $n_t$ the observed number of individuals with total score $t$, and with $p_{x|t}$ the proportion of these individuals who have profile $x$. Of course for $t = 0$ and $t = m$ there is no contribution to Equation 18, so we can also sum for $t = 1$ to $t = m - 1$. For the second simplification we write the second term of Equation 17 a bit more explicitly:

$$L_P(\varepsilon,F) = \sum_{t=0}^{m} \sum_{\substack{i \\ t_i=t}} \ln \int \pi_{\varepsilon}(t|\xi)\, dF_i(\xi). \tag{19}$$

Now $\pi_{i\varepsilon}(t_i)$ is linear in $F$, therefore $\ln \pi_{i\varepsilon}(t_i)$ is concave in $F_i$. Thus, given any $\varepsilon$, it follows that

$$\sum_{\substack{i \\ t_i=t}} \ln \int \pi_{\varepsilon}(t|\xi)\, dF_i(\xi) \leq n_t \ln \int \pi_{\varepsilon}(t|\xi)\, dF_t(\xi) \tag{20}$$

with

$$F_t(\xi) = n_t^{-1} \sum_{\substack{i \\ t_i=t}} F_i(\xi). \tag{21}$$

Since this inequality is particularly true at the maximum likelihood solution, we may suppose as well that all individual distributions $\{F_i|t_i = t\}$ are equal.

Thus, as far as maximization of Equation 17 is concerned, we may replace Equation 19 with

$$L_P(\varepsilon,F) = \sum_{t=0}^{m} n_t \ln \pi_{F_t,\varepsilon}(t). \tag{22}$$

It follows directly from this representation that we can never hope to estimate the individual $F_i$, only their averages $F_t$ by the method of maximum likelihood.

Combining Equations 18 and 22, we see that $L_T(\varepsilon,F)$ is like a product multinomial loglikelihood, paraphrasing theorem 2 of Cressie and Holland (1983) that this representation is a necessary condition for the Rasch model. Notice, however, that our result is more general: It applies to the functional as well as to the structural Rasch model. 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 Equation 12 that the optimum $F_i$ are one-point distributions. This also follows for the optimal $F_t$ from Equations 20 and 22. 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 or her ability $\xi_i$, 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 estimation in the classical functional Rasch model.

These unrestricted (also called *unconditional*) maximum likelihood estimates in the Rasch model have well-known problems. Andersen (1973) has shown that item parameter estimates 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 Andersen (1973), following suggestions of Rasch, has suggested the conditional maximum likelihood estimates of $\varepsilon$, 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 the latter interpretation (Andersen & Madsen, 1977), we complete the estimation process by maximizing $L_P(\hat{\varepsilon},F)$ over $\{F_t\}$ with $\hat{\varepsilon}$ the conditional maximum likelihood estimates of $\varepsilon$. Again, the optimum $F_t$ is one-step, where $\hat{\xi}_t$ solves the equation

$$\sum_{j=1}^{m} \xi\hat{\varepsilon}_j / (1 + \xi\hat{\varepsilon}_j) = t. \tag{23}$$

The conditional maximum likelihood estimates are consistent in the model $n \rightarrow \infty$ 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 Equation 23, 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* (Kiefer & Wolfowitz, 1956; Neyman & Scott, 1948). They cause inconsistencies and other kinds 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 is 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 $\varepsilon$ 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

$$L_C(\varepsilon) + L_P(\pi) \text{ is maximized, where}$$

$$L_P(\pi) = \sum_{t=0} n_t \ln \pi_t, \tag{24}$$

and the $\pi_t$ are unrestricted (except for the fact that they must be nonnegative numbers adding up to unity). The extended model is considered 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" (p. 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 (p. 137). We show that the MML estimates of $\varepsilon$ 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 CML 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 the 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 $\pi_{F,\hat{\varepsilon}}(t) = p_t$, 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 MML estimates, and $F$ is also maximum likelihood.

Let us analyze these equations a bit more in detail. They are

$$\gamma_t(\hat{\varepsilon})\int \xi^t \prod_{j=1}^{m} (1+\xi\hat{\varepsilon}_j)^{-1}\, dF(\xi) = p_t. \tag{25}$$

When is Equation 25 solvable for $F$? The functions that are integrated in Equation 25 are of the form $\alpha(t)\beta(\xi)\xi^t$. Thus they are a simple rescaling of the monomials $\xi^t$, and consequently they form a Tchebycheff system (Karlin & Studden, 1966, chap. I; Krein & Nudel'man, 1977, chap. II). Thus Equation 25 defines a generalized moment problem (or Tchebycheff moment problem) on the nonnegative real line. Such problems are analyzed in detail in Karlin and Studden (chap. V) and Krein and Nudel'man (chap. V). We follow Krein and Nudel'man, and first reduce Equation 25 in a simple way to a power moment problem (or Stieltjes moment problem). In the first place, Equation 25 is solvable if and only if

$$\int \varepsilon^t \prod_{j=1}^{m} (1+\xi\hat{\varepsilon})^{-1}\, dF(\xi) = p_t/\gamma_t(\hat{\varepsilon}) \tag{26}$$

is solvable.

Now since $dF(\xi)$ is a bounded measure on the positive half-line, so is

$$dG(\xi) = \prod_j (1+\xi\varepsilon_j)^{-1}\, dF(\xi)$$

so that Equation 26 is solvable if and only if

$$\int \xi^t\, dG(\xi) = p_t/\gamma_t(\hat{\varepsilon}) \tag{27}$$

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 & Studden, 1966, chap. V, section 10, or Krein & Nudel'man, 1977, chap. V, pp. 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 recapitulate 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 $B(\hat{\varepsilon})$ 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 $b_{st}(\hat{\varepsilon}) = p_{s+t+1}/\gamma_{s+t+1}(\hat{\varepsilon})$ for $s, t = 0, \ldots, \frac{1}{2}(m-1)$. The power moment problem (Equation 27) is solvable for $G$ if and only if $A(\hat{\varepsilon})$ and $B(\hat{\varepsilon})$ are both positive semidefinite. If they are both positive definite, then a solution without mass at infinity exists. If a

solution 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 MML estimates of the $\varepsilon_j$ are equal to the CML 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 MML estimates are equal to CML estimates with probability tending to one. This implies, of course, that they have the same asymptotic normal distribution, and they are both efficient in the structural Rasch model. This constitutes our main result, because it shows that the extended Rasch model of Tjur and Cressie and 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 MML may become quite different from CML. 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, chap. V, section 4). For the Stieltjes power moment problem, even more details are given by Krein and Nudel'man (1977, chap. 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 $u_t = \xi^t \prod_{j=1}^{m} (1 + \xi\varepsilon_j)^{-1}$, $t = 0, 1, \ldots, m$, 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 $\xi$ converges to infinity. We are only interested in the lower principal representation, because the upper one places mass at infinity. This lower principal 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 $F$ is estimated consistently. But there is no reason to believe that $F$ has this highly artificial form in real life 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 may 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, some of which are certainly relevant for further study of the Rasch model.

We briefly recapitulate the major result of this section. 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 MML estimates of the item parameters are equal to the CML 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, it 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 nonstandard 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.

### *An Example*

Computing CML estimates is generally not an easy task. Compare Wainer, Morgan, and Gustafsson (1980) and Gustafsson (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 that are asymtotically equivalent to them, by solving a *finite mixture problem* (Everitt & Hand, 1981; Redner & 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

$$L_t(\varepsilon,\xi,\theta) = \sum_{j=1}^{m} s_j \ln \varepsilon_j + \sum_{t=0}^{m} n_t \ln \sum_{\nu}^{r} \theta_\nu \xi_\nu^t \prod_{j=1}^{m} (1 + \xi_\nu \varepsilon_j)^{-1}, \tag{28}$$

where $r = \begin{cases} (m+1)/2 & \text{if } m \text{ is odd} \\ (m+2)/2 & \text{if } m \text{ is even.} \end{cases}$

In the case where $m$ is even, one knot ($\xi$-value with a jump in the distribution function) is set equal to zero. The solution can be normalized by setting one (free) knot equal to one, and with the additional constraint that the masses $\theta_\nu$ must add up to one, the number of free parameters in the model is given by

$m + \frac{1}{2}(m-1) + \frac{1}{2}(m-1)$ if $m$ is odd, and $m + (m/2 - 1) + m/2$ if $m$ is even, yielding $2m - 1$ free parameters in both cases. Since there are $2^m - 1$ free profile probabilities, the number of degrees of freedom for testing the model is $2^m - 2m$. (Note that our recourse to the lower principal representation gives the number of degrees of freedom in a fairly direct way, whereas Cressie and Holland need a rather subtle argument to arrive at the same conclusion.)

To get an impression of the canonical representation of $F$ we analyzed the responses of 1,000 subjects to a subset of 5 items from the Law School Admissions Test, referred to as LSAT-6 (cfr. Mislevy, 1984, or Thissen, 1982 for the raw data). The algorithmic problems we encountered, especially the very slow convergence of the *EM* algorithm, are not essential here and will be reported elsewhere.

The results can be summarized as follows: the item parameter estimates $\hat{\varepsilon}$ are exactly the same as the CML estimates, the matrices $A(\hat{\varepsilon})$ and $B(\hat{\varepsilon})$ being positive definite. The lower principal representation of $F$ for the LSAT-6 data is given below.

| $\hat{\xi}_\nu$ | $\hat{\theta}_\omega$ |
|---|---|
| 0.5896 | 0.04035 |
| 3.2366 | 0.74199 |
| 21.397 | 0.21766 |

It should be noted that although this representation is of little practical value, it allows the loglikelihood function to take its absolute maximum without any restriction put on the distribution of the latent variable, except that it be a true nondegenerate distribution. Its value in this case is $-2466.47$, whereas the likelihood of the (saturated) multinomial model for the profile probabilities is $-2456.04$, giving rise to a likelihood ratio statistic of 20.86, which with $32 - 10 = 22$ degrees of freedom will certainly not lead to rejection of the Rasch model.

## Conclusion

In this paper we considered MML estimation of a general, structural Rasch model. In this model it is assumed that each individual has its own ability distribution $F_i$. It is pointed out that maximum likelihood estimates always attribute the same distribution $F_t$ to all individuals having total score $t$. If no restrictions are put on the $F_i$, their maximum likelihood estimates are one-point distributions. In this case the maximum likelihood estimates are equivalent to the unconditional maximum likelihood estimates in the functional model, where each individual is characterized by his or her ability $\xi_i$. Several restrictions may be put on the $F_i$: They may all be thought identical and/or they may be parametrized in several ways.

We investigated the special case where $F_i = F$, with the only restriction that $F$ be a nondegenerate distribution function. It was shown that in this case the maximum likelihood estimates of the item parameters equal the conditional maximum likelihood estimates, with probability tending to one as $n \to \infty$. A unique canonical form of $F$ can be estimated together with the item parameters by solving a finite mixture problem, which avoids the computationally hard task of computing the CML estimates directly.

This model was also shown to be asymptotically equivalent to the extended Rasch model discussed by Tjur and by Cressie and Holland.

## 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.

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.

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

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions.* London: Chapman and Hall.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Huber.

Gustafsson, J. E. (1979). *PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items* (Rep. No. 85). Göteborg: Institute of Education.

Gustafsson, 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, RI: 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 increase without limit.* Princeton, NJ: Educational Testing Service.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: 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. Reissued, 1980, Chicago, University of Chicago Press.

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

Rasch, G. (1966). An informal report on a theory of objectivity in comparisons. In L. J. Th. van der Kamp & C. A. J. Vlek (Eds.), *Proceedings of the NUFFIC 1966 Summer Session* (pp. 181–199). The Hague, Netherlands: NUFFIC.

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 responses. *Journal of the American Statistical Association, 69,* 744–749.

Sanathanan, L., & Blumenthal, 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, 1,* 245–262.

Wainer, H., Morgan, A., & Gustafsson, 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.

## Authors

JAN DE LEEUW, Professor, Department of Data Theory, State University of Leiden, Middelstegracht 4, 2312 TW Leiden, The Netherlands. *Specialization:* Multivariate analysis.

NORMAN VERHELST, National Institute of Educational Measurement (CITO), P.O. Box 1034, 6801 Arnhem, The Netherlands. *Specialization:* Latent trait theory.