

Minorization Algorithms for the Rasch Model

Gunter Maris
Timo M. Bechger



Minorization Algorithms for the Rasch Model

Gunter Maris

Timo M. Bechger

Cito
Arnhem, 2009

This manuscript has been submitted for publication. No part of this manuscript may be copied or reproduced without permission.

Abstract

From the principle of minorization, we derive monotone convergent algorithms for conditional and marginal maximum likelihood estimation in the Rasch model, where the parameters are updated one item at the time. In addition, we show that these algorithms can be made faster by deliberately over-parameterizing the model. The algorithm for CML turns out to be equal to the implicit equations algorithm that was proposed by Gerhard Fischer in the early 1970s, without proof of its monotone convergence.

In this paper we consider estimation methods for the parameters in the Rasch (1960) model based on the idea of minorization (De Leeuw, 1994). Minorization provides a general framework for constructing monotone convergent algorithms for parameter estimation. The iterative minorization approach for finding the maximum of a (log-likelihood) function $l(\epsilon_j)$ (De Leeuw, 1994; Heiser, 1995; Lange, Hunter, & Yang, 2000) rests on the following chain of inequalities:

$$l(\epsilon_j) \geq M(\epsilon_j, \hat{\epsilon}) > M(\hat{\epsilon}_j, \hat{\epsilon}) = l(\hat{\epsilon}_j)$$

known as the *sandwich* inequality, where the auxiliary function M is called a *minorizing* function. From the sandwich inequality we obtain the following

properties for the minorizing function M :

$$l(\epsilon_j) \geq M(\epsilon_j, \hat{\boldsymbol{\epsilon}}) \quad \text{and} \quad M(\hat{\epsilon}_j, \hat{\boldsymbol{\epsilon}}) = l(\hat{\epsilon}_j) \quad (1)$$

Repeatedly constructing and maximizing a minorizing function yields an algorithm that will in every step increase the value of the likelihood: $M(\epsilon_j, \hat{\boldsymbol{\epsilon}}) > M(\hat{\epsilon}_j, \hat{\boldsymbol{\epsilon}}) \Rightarrow l(\epsilon_j) > l(\hat{\epsilon}_j)$. Since the first order derivatives of the minorizing function and the loglikelihood are equal when evaluated at $\hat{\boldsymbol{\epsilon}}$ we obtain that:

$$\left. \frac{\partial}{\partial \epsilon_j} M(\epsilon_j, \hat{\boldsymbol{\epsilon}}) \right|_{\epsilon_j = \hat{\epsilon}_j} = 0 \Leftrightarrow \left. \frac{\partial}{\partial \epsilon_j} l(\epsilon_j) \right|_{\epsilon_j = \hat{\epsilon}_j} = 0 \quad (2)$$

Hence, if iterative maximization of the minorizing function converges (i.e., the left hand side of Equation 2), we have found the parameters values at which the function $l(\epsilon_j)$ reaches its maximum (i.e., the right hand side of Equation 2).

We derive algorithms for conditional and marginal maximum likelihood. The parameters are updated one item at the time, and the update for the j th item parameter is of the form

$$\epsilon_j = \hat{\epsilon}_j \frac{O_j}{\hat{\mathcal{E}}_j} \quad (3)$$

where O_j and $\hat{\mathcal{E}}_j$, evaluated at $\hat{\epsilon}_j$, denote the observed and expected value of the sufficient statistic for ϵ_j . The difference between conditional and marginal maximum likelihood is in the expected value $\hat{\mathcal{E}}_j$. By deliberately over-parameterizing the model, a faster algorithm is derived where the update factor is the ratio of observed and expected *odds-ratios*. An example is used to illustrate how these algorithms generalize to more complex IRT models.

The paper is organized as follows. In the first section, we introduce the Rasch model and consider the different estimation methods. In the second section, a minorization algorithm for computing conditional maximum likelihood estimates is considered. In the third section, a similar minorization algorithm is considered for computing marginal maximum likelihood estimates. The fourth section deals with ways to accelerate convergence. In the fifth section, we show how the minorization approach generalizes to more complex IRT models. The paper ends with a discussion.

1 The Rasch model

The Rasch model (Rasch, 1960) is defined as follows:

$$P(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\delta}) = \prod_{p=1}^N \prod_{i=1}^M \frac{\exp(x_{pi} [\theta_p - \delta_i])}{1 + \exp(\theta_p - \delta_i)} \quad (4)$$

where the responses of N persons to M items are represented by a matrix \mathbf{x} where the rows are response patterns $\mathbf{x}_p = (x_{p1}, \dots, x_{pM})$. The first item parameter δ_1 is arbitrarily set to zero to identify the model and the other item parameters are interpreted as difficulty relative to the first item. A recent and extensive survey of the literature related to the Rasch model can be found in Fischer (2007).

If we wish to estimate the item parameters δ_i , the person parameters θ_p are *incidental*. That is, their number increases with the sample size. It is known that in the presence of an increasing number of incidental parameters it is, in general, not possible to estimate the (structural) item parameters consistently (Neyman & Scott, 1948). This problem can be overcome in one of

two ways. First, since the Rasch model is an exponential family model (e.g., Andersen, 1977) we can base our inferences on the distribution of the data conditionally on the sufficient statistics for the incidental parameters. This method is called *conditional maximum likelihood* (CML) estimation. Second, if the subjects can be regarded as a random sample from a well defined population characterized by an ability distribution G , the structural item parameters can be estimated from the marginal distribution of the data. That is, we integrate the incidental parameters out of the model. Rather than estimating each subject's ability, only the parameters of the ability *distribution* need to be estimated. This method is called *marginal maximum likelihood* (MML) estimation. Under suitable regularity conditions both methods can be shown to lead to consistent estimates of the item difficulty parameters.

The likelihood of the Rasch model conditional upon the sum scores:

$$X_{p+} = \sum_i X_{pi}$$

has the following form:

$$P(\mathbf{x}|\mathbf{x}_+, \boldsymbol{\delta}) = \frac{\prod_i \epsilon_i^{x_{+i}}}{\prod_s \gamma_s^{n_s}(\boldsymbol{\epsilon})}$$

where $\mathbf{x}_+ = (x_{1+}, \dots, x_{N+})$, $\epsilon_i \equiv \exp(-\delta_i)$, n_s denotes the number of persons with $x_{p+} = s$, and the $\gamma_j(\boldsymbol{\epsilon})$ (or γ_j for short) are the *elementary symmetric functions* (e.g., Hardy, Littlewood, & Polya, 1952, 2.22, Andersen, 1972; Verhelst, Glas, & van der Sluis, 1984). It is customary to find the maximum of the logarithm of the likelihood:

$$l_{\mathbf{x}|\mathbf{x}_+}(\boldsymbol{\epsilon}) = \sum_i x_{+i} \ln(\epsilon_i) - \sum_s n_s \ln(\gamma_s) \quad (5)$$

by setting its derivatives equal to zero:

$$\frac{x_{+j}}{\epsilon_j} - \sum_s n_s \frac{\gamma_{s-1}^{(j)}}{\gamma_s} = 0 \quad (6)$$

where $\gamma_s^{(j)} = \frac{\partial}{\partial \epsilon_j} \gamma_s$ denotes the elementary symmetric function of $\boldsymbol{\epsilon}$ without ϵ_j (Andersen, 1972). This expression can be cast in the general form of exponential family likelihood equations (Fischer, 1995; Andersen, 1980):

$$x_{+j} = \sum_s n_s \frac{\epsilon_j \gamma_{s-1}^{(j)}}{\gamma_s} = \mathcal{E}(X_{+j} | \mathbf{x}_{.+}, \boldsymbol{\epsilon}) \quad (7)$$

Observe that the likelihood equations need not always have a finite solution. Fischer (1981), Haberman (1977) show that a finite solution exists if and only if the data are *well conditioned*. Throughout this article we assume well conditioned data.

The marginal Rasch model assumes that abilities have been sampled from a population distribution $g(\tau)$, where $\tau \equiv \exp(\theta)$.¹ The marginal log-likelihood is

$$\ln P(\mathbf{x}) = \sum_p \ln \int_R \prod_i \frac{\tau^{x_{pi}} \epsilon_i^{x_{pi}}}{1 + \tau \epsilon_i} g(\tau) d\tau \quad (8)$$

with derivatives

$$\frac{\partial}{\partial \epsilon_j} \ln P(\mathbf{x}) = \sum_p \int_R \left[\frac{\partial}{\partial \epsilon_j} \ln P(x_{pj} | \tau, \epsilon_j) \right] f(\tau | \mathbf{x}_p; \boldsymbol{\epsilon}) d\tau \quad (9)$$

$$= \sum_p \int_R \left[\frac{x_{pj}}{\epsilon_j} - \frac{\tau}{1 + \tau \epsilon_j} \right] f(\tau | \mathbf{x}_p; \boldsymbol{\epsilon}) d\tau \quad (10)$$

¹The parameters of the population distribution are ignored for simplicity of presentation.

where $f(\tau|\mathbf{x}_p; \boldsymbol{\epsilon})$ is the posterior distribution of ability τ . Equating this derivative with zero gives the marginal likelihood equation:

$$\frac{x_{+j}}{\epsilon_j} = \sum_p \int_R \frac{\tau}{1 + \tau\epsilon_j} f(\tau|\mathbf{x}_p; \boldsymbol{\epsilon}) d\tau \quad (11)$$

Note that this expression can be written as

$$x_{+j} = \sum_p \int_R \frac{\tau\epsilon_j}{1 + \tau\epsilon_j} f(\tau|\mathbf{x}_p; \boldsymbol{\epsilon}) d\tau = \mathcal{E}[\mathcal{E}(X_{+j}|\tau) | \mathbf{x}; \boldsymbol{\epsilon}] \quad (12)$$

Even though the marginal likelihood is not a member of the exponential family, the CML (7) and MML likelihood equations (12) are quite similar. Neither the CML nor the MML likelihood equations admit an analytical solution, and iterative methods have to be used to find a solution.

2 A Minorization Algorithm for CML

The second term in the conditional log-likelihood in Equation 5 is intractable because it contains logarithms of elementary symmetric functions. Hence, we develop a minorization for these functions.

The elementary symmetric functions satisfy the following recursive relation:

$$\gamma_s = \gamma_s^{(j)} + \epsilon_j \gamma_{s-1}^{(j)}, \quad (\gamma_0^{(j)} = 1, \text{ and } \gamma_s^{(j)} = 0 \text{ if } s < 0 \text{ or } s > k - 1)$$

(Andersen, 1972; Verhelst et al., 1984). Hence, the conditional log-likelihood function can be written as:

$$l_{\mathbf{x}|\mathbf{x}_+}(\boldsymbol{\epsilon}) = \sum_i x_{+i} \ln(\epsilon_i) - \sum_s n_s \ln \left(\gamma_s^{(j)} + \epsilon_j \gamma_{s-1}^{(j)} \right)$$

and we find that the complicated part in the conditional log-likelihood is actually rather simple when considered as a function of a single item parameter.²

The terms $\ln(\gamma_s^{(j)} + \epsilon_j \gamma_{s-1}^{(j)})$ are concave functions in ϵ_j . A concave function lies below any of its tangents so that:

$$\ln\left(\gamma_s^{(j)} + \epsilon_j \gamma_{s-1}^{(j)}\right) \leq \ln\left(\gamma_s^{(j)} + \widehat{\epsilon}_j \gamma_{s-1}^{(j)}\right) + \frac{\gamma_{s-1}^{(j)}}{\gamma_s^{(j)} + \widehat{\epsilon}_j \gamma_{s-1}^{(j)}} (\epsilon_j - \widehat{\epsilon}_j) \quad (13)$$

with equality when $\epsilon_j = \widehat{\epsilon}_j$. Substituting the right hand side of (13) for the left hand side in the conditional log-likelihood we obtain the following minorizing function:

$$\begin{aligned} l_{\mathbf{x}|\mathbf{x}_+}(\boldsymbol{\epsilon}) &\geq \sum_i x_{+i} \ln(\epsilon_i) - \sum_s n_s \left(\ln(\gamma_s^{(j)} + \widehat{\epsilon}_j \gamma_{s-1}^{(j)}) + \frac{\gamma_{s-1}^{(j)}}{\gamma_s^{(j)} + \widehat{\epsilon}_j \gamma_{s-1}^{(j)}} (\epsilon_j - \widehat{\epsilon}_j) \right) \\ &= M(\epsilon_j, \widehat{\boldsymbol{\epsilon}}). \end{aligned} \quad (14)$$

The derivative of $M(\epsilon_j, \widehat{\boldsymbol{\epsilon}})$ with respect to ϵ_j is:

$$\frac{x_{+j}}{\epsilon_j} - \sum_s n_s \frac{\gamma_{s-1}^{(j)}}{\gamma_s^{(j)} + \widehat{\epsilon}_j \gamma_{s-1}^{(j)}}.$$

If we compare this with the derivative, in Equation 6, of the conditional log-likelihood we see that the only difference is that the denominator in the second term now depends on $\widehat{\epsilon}_j$ alone. Setting the derivative of the minorizing function equal to zero, we see that we can find an explicit solution for ϵ_j :

$$\epsilon_j = \frac{x_{+j}}{\sum_s n_s \frac{\gamma_{s-1}^{(j)}}{\gamma_s^{(j)} + \widehat{\epsilon}_j \gamma_{s-1}^{(j)}}}$$

² $\gamma^{(j)}$ does not depend on ϵ_j .

If we multiply the right hand side numerator and denominator by $\widehat{\epsilon}_j$ we obtain the following more compact form:

$$\epsilon_j = \widehat{\epsilon}_j \frac{x_{+j}}{\sum_s n_s \frac{\gamma_{s-1}^{(j)} \widehat{\epsilon}_j}{\gamma_s^{(j)} + \widehat{\epsilon}_j \gamma_{s-1}^{(j)}}} = \widehat{\epsilon}_j \frac{x_{+j}}{\mathcal{E}(X_{+j} | \mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})} \quad (15)$$

where $\mathcal{E}(X_{+j} | \mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})$ denotes the conditional expectation of X_{+j} , evaluated at $\boldsymbol{\epsilon} = \widehat{\boldsymbol{\epsilon}}$. In the psychometric literature, this algorithm is known as the *implicit equation algorithm*. It was proposed by Fischer (1974) without proof of its (monotone) convergence. For later reference we state this as

Theorem 1. *Fischer's implicit equation algorithm is an iterative minorization algorithm, and therefore monotonely convergent.*

3 Minorization Algorithms for MML

In order to derive a minorizing function for the marginal likelihood function, it turns out to be convenient to consider the logarithm of a likelihood ratio:

$$\begin{aligned} \ln \frac{P(\mathbf{x}_p; \boldsymbol{\epsilon})}{P(\mathbf{x}_p; \widehat{\boldsymbol{\epsilon}})} &= \sum_p \ln \int_R \frac{P(\mathbf{x}_p | \tau; \boldsymbol{\epsilon}) g(\tau)}{P(\mathbf{x}_p; \widehat{\boldsymbol{\epsilon}})} d\tau \\ &= \sum_p \ln \int_R \frac{P(\mathbf{x}_p | \tau; \boldsymbol{\epsilon})}{P(\mathbf{x}_p | \tau; \widehat{\boldsymbol{\epsilon}})} \frac{P(\mathbf{x}_p | \tau; \widehat{\boldsymbol{\epsilon}}) g(\tau)}{P(\mathbf{x}_p; \widehat{\boldsymbol{\epsilon}})} d\tau \\ &= \sum_p \ln \int_R \frac{P(\mathbf{x}_p | \tau; \boldsymbol{\epsilon})}{P(\mathbf{x}_p | \tau; \widehat{\boldsymbol{\epsilon}})} f(\tau | \mathbf{x}_p; \widehat{\boldsymbol{\epsilon}}) d\tau \\ &= \sum_p \ln E \left[\frac{P(\mathbf{x}_p | \tau; \boldsymbol{\epsilon})}{P(\mathbf{x}_p | \tau; \widehat{\boldsymbol{\epsilon}})} \middle| \mathbf{x}_p; \widehat{\boldsymbol{\epsilon}} \right] \end{aligned} \quad (16)$$

Clearly, finding the parameter values at which this log likelihood ratio reaches its largest value is equivalent to finding the parameter values at which the log-likelihood reaches its largest value.

3.1 The EM-algorithm

Since $\ln(x)$ is a strictly concave function on $(0, \infty)$, it follows from Jensen's inequality that

$$\begin{aligned} \ln \frac{P(\mathbf{x}_p; \boldsymbol{\epsilon})}{P(\mathbf{x}_p; \hat{\boldsymbol{\epsilon}})} &= \sum_p \ln \left(E \left[\frac{P(\mathbf{x}_p|\tau; \boldsymbol{\epsilon})}{P(\mathbf{x}_p|\tau; \hat{\boldsymbol{\epsilon}})} \middle| \mathbf{x}_p; \hat{\boldsymbol{\epsilon}} \right] \right) \\ &\geq \sum_p E \left[\ln \left(\frac{P(\mathbf{x}_p|\tau; \boldsymbol{\epsilon})}{P(\mathbf{x}_p|\tau; \hat{\boldsymbol{\epsilon}})} \right) \middle| \mathbf{x}_p; \hat{\boldsymbol{\epsilon}} \right] \\ &\geq \sum_p E [\ln (P(\mathbf{x}_p|\tau; \boldsymbol{\epsilon})) | \mathbf{x}_p; \hat{\boldsymbol{\epsilon}}] - \sum_p E [\ln (P(\mathbf{x}_p|\tau; \hat{\boldsymbol{\epsilon}})) | \mathbf{x}_p; \hat{\boldsymbol{\epsilon}}] \end{aligned} \quad (17)$$

with equality when $\boldsymbol{\epsilon} = \hat{\boldsymbol{\epsilon}}$. If we expand the function on the right side of (17), we see that we have found a minorizing function

$$M(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}}) = Q(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}}) - Q(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}}) \quad (18)$$

where

$$Q(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}}) = \sum_p E [\ln P(\mathbf{x}_p|\tau; \boldsymbol{\epsilon}) | \mathbf{x}_p; \hat{\boldsymbol{\epsilon}}] \quad (19)$$

It is seen from (18) that $M(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}})$ improves if $Q(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}})$ improves; $Q(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}})$ is constant. An iterative minorization algorithm where, in each step, the Q -function is maximized is called an *EM-algorithm* (Dempster, Laird, & Rubin, 1977; de Leeuw, 2006).

Under the Rasch model,

$$Q(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}}) = \sum_p \sum_i \int_R (x_{pi} \ln \epsilon_i - \ln(1 + \tau \epsilon_i)) f(\tau | \mathbf{x}_p; \hat{\boldsymbol{\epsilon}}) d\tau. \quad (20)$$

Setting the derivative of $M(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}})$ with respect to ϵ_j to zero gives

$$\frac{x_{+j}}{\epsilon_j} = \sum_p \int_R \frac{\tau}{1 + \tau \epsilon_j} f(\tau | \mathbf{x}_p; \hat{\boldsymbol{\epsilon}}) d\tau \quad (21)$$

The MLE is obtained by solving this equation for ϵ_j . Observe that, in comparison with the marginal likelihood equation (11), the posterior distribution of τ no longer depends on ϵ but only on its current value. We see that with an EM algorithm we can update the item parameters one at a time, since Equation 21 only depends on a single item parameter.

Maximization of the minorizing function M is easier than maximization of the full marginal likelihood. However, no closed form for the maximum has been obtained, and the solution still requires a numerical optimization method.

3.2 Minorizing the Q -function of EM

Reconsidering the Q -function, we recognize the term $\ln(1 + \tau\epsilon_i)$ as a concave function of ϵ_i . Hence,

$$-\ln(1 + \tau\epsilon_i) \geq -\ln(1 + \tau\hat{\epsilon}_i) - \frac{\tau}{1 + \tau\hat{\epsilon}_i}(\epsilon_i - \hat{\epsilon}_i)$$

Using this inequality in (20) gives that, for all $\tau > 0$,

$$\ln P(x_{pi}|\tau; \epsilon_i) f(\tau|\mathbf{x}_p; \hat{\epsilon}) \geq \left(x_{pi} \ln \epsilon_i - \ln(1 + \tau\hat{\epsilon}_i) - \frac{\tau}{1 + \tau\hat{\epsilon}_i}(\epsilon_i - \hat{\epsilon}_i) \right) f(\tau|\mathbf{x}_p; \hat{\epsilon})$$

with equality when $\epsilon_i = \hat{\epsilon}_i$. Hence we may minorize the Q -function as follows:

$$\begin{aligned} Q^*(\epsilon, \hat{\epsilon}) &= \sum_i x_{+i} \ln \epsilon_i - \sum_p \sum_i \int_R \left(\ln(1 + \tau\hat{\epsilon}_i) + \frac{\tau}{1 + \tau\hat{\epsilon}_i}(\epsilon_i - \hat{\epsilon}_i) \right) f(\tau|\mathbf{x}_p; \hat{\epsilon}) d\tau \\ &\leq Q(\epsilon, \hat{\epsilon}) \end{aligned}$$

Consequently, we have found an alternative minorizing function for the marginal log likelihood ratio:

$$M^*(\epsilon, \hat{\epsilon}) = Q^*(\epsilon, \hat{\epsilon}) - Q(\hat{\epsilon}, \hat{\epsilon}) \leq M(\epsilon, \hat{\epsilon}) \leq \ln \frac{P(\mathbf{x}_p; \epsilon)}{P(\mathbf{x}_p; \hat{\epsilon})}$$

This gives rise to an alternative algorithm. Setting the derivative of $M^*(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}})$ with respect to ϵ_j equal to zero gives

$$\frac{x_{+j}}{\epsilon_j} = \sum_p \int_R \frac{\tau}{1 + \tau \hat{\epsilon}_j} f(\tau | \mathbf{x}_p; \hat{\boldsymbol{\epsilon}}) d\tau \quad (22)$$

In comparison to Equation 21, the integral on the right hand side is now completely independent of ϵ_j , and we can solve Equation 22 explicitly for ϵ_j :

$$\epsilon_j = \hat{\epsilon}_j \frac{x_{+j}}{\sum_p \int_R \frac{\tau \hat{\epsilon}_j}{1 + \tau \hat{\epsilon}_j} f(\tau | \mathbf{x}_p; \hat{\boldsymbol{\epsilon}}) d\tau} = \hat{\epsilon}_j \frac{x_{+j}}{\mathcal{E}[\mathcal{E}(X_{+j} | \tau, \hat{\epsilon}_j) | \mathbf{x}; \hat{\boldsymbol{\epsilon}}]} \quad (23)$$

Comparing this algorithm for marginal maximum likelihood estimation to the one for conditional maximum likelihood estimation in Equation 15 we see that both algorithms are of the same form (3) and differ only in the expected value of the sufficient statistic that is used. Although this algorithm for MML requires more iterations than the EM algorithm, it is considerably simpler because all terms on the right side of (22) are evaluated at the current estimates.

4 Overparameterization

If the data \mathbf{x} are recoded such that $y_{pi} = 1 - x_{pi}$, the likelihood for the recoded data \mathbf{y} still conforms to the Rasch model. Specifically,

$$\begin{aligned} p(\mathbf{y} | \boldsymbol{\theta}^*, \boldsymbol{\delta}^*) &= \prod_p \prod_i \frac{\exp((1 - x_{pi})(\theta_p^* - \delta_i^*))}{1 + \exp((\theta_p^* - \delta_i^*))} \\ &= \prod_p \prod_i \frac{\exp(x_{pi}([- \theta_p^*] - [- \delta_i^*]))}{1 + \exp([- \theta_p^*] - [- \delta_i^*])} \end{aligned}$$

with parameters $\theta_p^* \equiv -\theta_p$ and $\delta_i^* \equiv -\delta_i$. The sufficient statistics for these new parameters are $y_{p+} = M - x_{p+}$ and $y_{+i} = N - x_{+i}$ respectively. Consequently, the algorithms derived in this paper can be used for the original

data \mathbf{x} as well as for the recoded data \mathbf{y} . However, the parameter updates for \mathbf{x} and \mathbf{y} are *not* equivalent. Specifically,

$$\begin{aligned} \exp(-\delta_i^*) &= \exp(-\widehat{\delta}_i^*) \frac{y_{+i}}{\mathcal{E}(Y_{+i}|\mathbf{y}_{.+}, \widehat{\boldsymbol{\epsilon}})} = \exp(-\widehat{\delta}_i^*) \frac{N - x_{+i}}{N - \mathcal{E}(X_{+i}|\mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})} \\ &\Downarrow \\ \exp(-\delta_i) &= \exp(-\widehat{\delta}_i) \frac{N - \mathcal{E}(X_{+i}|\mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})}{N - x_{+i}} \end{aligned}$$

The algorithms differ in the update factor. Both updates increase the log-likelihood, and we can determine which takes the largest step

$$\frac{x_{+i}}{\mathcal{E}(X_{+i}|\mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})} \quad \text{or} \quad \frac{N - \mathcal{E}(X_{+i}|\mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})}{N - x_{+i}}$$

Next we show that for both of these update factors, the new value is between the current value and the value at which the conditional log-likelihood attains its largest value. Hence, we may choose at any point that update factor that gives the greater change in parameter value, knowing that it will also lead to the higher likelihood value.

We show that if we denote the value of ϵ_j at which the conditional log-likelihood, considered as a function of the single parameter ϵ_j , attains its maximum by $\bar{\epsilon}_j$, the implicit equation algorithm leads to a new value between $\widehat{\epsilon}_j$ and $\bar{\epsilon}_j$. We prove this by showing that the sign of the derivative of the conditional log-likelihood with respect to ϵ_j has not changed when the parameter has been updated with the implicit equation algorithm.

First, if we assume that $\widehat{\epsilon}_j < \bar{\epsilon}_j$, we easily find that the derivative of the conditional log-likelihood with respect to ϵ_j is positive:

$$[\widehat{\epsilon}_j < \bar{\epsilon}_j] \Rightarrow [\widehat{\mathcal{E}}(X_{+j}) < \bar{\mathcal{E}}(X_{+j}) = x_{+j}] \Rightarrow \left[\frac{\partial}{\partial \epsilon_j} l_{\mathbf{x}|\mathbf{x}_{.+}}(\boldsymbol{\epsilon}) \Big|_{\epsilon_j = \widehat{\epsilon}_j} > 0 \right]$$

where $\mathcal{E}(X_{+j})$ is short-hand for $\mathcal{E}(X_{+j}|\mathbf{x}_{+}, \boldsymbol{\epsilon})$, and we readily see that the parameter value increases.

Second, filling in the new value in the derivative of the conditional log-likelihood with respect to ϵ_j (6) reveals that the sign of the derivative does not change:

$$\begin{aligned} \frac{\partial}{\partial \epsilon_j} l_{\mathbf{x}|\mathbf{x}_{+}}(\boldsymbol{\epsilon}) &= \frac{x_{+j}}{\epsilon_j} - \sum_p \frac{\gamma_{n_p-1}^{(j)}}{\gamma_{n_p}^{(j)} + \gamma_{n_p-1}^{(j)} \epsilon_j} \\ &\Downarrow \text{ fill in } \epsilon_j = \widehat{\epsilon}_j \widehat{\mathcal{E}}(X_{+j}) \text{ in the first term} \\ 0 &< \sum_p \frac{\gamma_{n_p-1}^{(j)}}{\gamma_{n_p}^{(j)} + \gamma_{n_p-1}^{(j)} \widehat{\epsilon}_j} - \sum_p \frac{\gamma_{n_p-1}^{(j)}}{\gamma_{n_p}^{(j)} + \gamma_{n_p-1}^{(j)} \epsilon_j} \end{aligned} \quad (24)$$

This means that, all other things being equal, we never overshoot the maximum value of the likelihood.

A different perspective on this curious result, and one that allows further progress, is obtained if we consider different ways to parameterize the Rasch model. Rather than working with the likelihood in (4) we may write the model as follows

$$\begin{aligned} P(\mathbf{x}|\boldsymbol{\tau}, \boldsymbol{\nu}) &= \prod_p \frac{\tau_p^{x_{p+}} \prod_i \nu_i^{1-x_{pi}}}{\prod_i (\nu_i + \tau_p)} \\ &= \prod_p \frac{\tau_p^{x_{p+}} \prod_i \nu_i^{1-x_{pi}}}{\sum_{j=0}^M \gamma_j(\mathbf{1}, \boldsymbol{\nu}) \tau_p^j} \end{aligned} \quad (25)$$

where obviously, $\nu_i = \epsilon_i^{-1}$. We readily find that $N - x_{+i}$ is sufficient for ν_i and a similar implicit equation algorithm can be derived for estimating the ν parameters from their conditional likelihood. We find that this implicit equation algorithm corresponds to the implicit equation algorithm obtained when the data are recoded.

Next we consider a situation where we use yet another parameterization of the Rasch model

$$\begin{aligned}
P(\mathbf{x}|\boldsymbol{\tau}, \boldsymbol{\epsilon}, \boldsymbol{\nu}) &= \prod_p \frac{\tau_p^{n_p} \prod_i \epsilon_i^{x_{pi}} \nu_i^{1-x_{pi}}}{\prod_i (\nu_i + \epsilon_i \tau_p)} \\
&= \prod_p \frac{\tau_p^{n_p} \prod_i \epsilon_i^{x_{pi}} \nu_i^{1-x_{pi}}}{\sum_{j=0}^M \gamma_j(\boldsymbol{\epsilon}, \boldsymbol{\nu}) \tau_p^j}
\end{aligned} \tag{26}$$

It is obvious that we cannot identify the parameters $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ from the resulting conditional likelihood. We can only identify the ratio of ϵ_i over ν_i . However, if a set of parameters is found which maximizes the likelihood in (26), we can easily deduce from it the maximum likelihood estimators assuming any set of identifying constraints. As we will show, it is computationally convenient to start from an *overparameterized* model in which the parameters are not identifiable.

It is readily found that the likelihood in Equation 26 gives rise to a conditional log-likelihood of the following form:

$$\ln P(\mathbf{x}|\mathbf{x}_{.+}, \boldsymbol{\epsilon}, \boldsymbol{\nu}) = \sum_p \sum_i x_{pi} \ln(\epsilon_i) + (1 - x_{pi}) \ln(\nu_i) - \sum_p \ln \gamma_{n_p}(\boldsymbol{\epsilon}, \boldsymbol{\nu})$$

The functions

$$\ln \gamma_{n_p}(\boldsymbol{\epsilon}, \boldsymbol{\nu}) = \ln \left(\gamma_{n_p}^{(i)}(\boldsymbol{\epsilon}, \boldsymbol{\nu}) \nu_i + \gamma_{n_p-1}^{(i)}(\boldsymbol{\epsilon}, \boldsymbol{\nu}) \epsilon_i \right)$$

are concave both in ϵ_i and in ν_i . Hence, a bivariate minorization is easily achieved, following a line of thought similar to the one that led to the minorization used for the derivation of the implicit equation algorithm.

Working out the details gives the following updates for ϵ_i and ν_i :

$$\epsilon_i = \hat{\epsilon}_i \frac{x_{+i}}{\sum_p \frac{\hat{\epsilon}_i \gamma_{x_{p+}-1}^{(i)}(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\hat{\nu}_i \gamma_{x_{p+}}^{(i)}(\boldsymbol{\epsilon}, \boldsymbol{\nu}) + \hat{\epsilon}_i \gamma_{x_{p+}-1}^{(i)}(\boldsymbol{\epsilon}, \boldsymbol{\nu})}} = \hat{\epsilon}_i \frac{x_{+i}}{\mathcal{E}(X_{+i}|\mathbf{x}_{.+}, \hat{\boldsymbol{\epsilon}})} \tag{27}$$

and

$$\nu_i = \widehat{\nu}_i \frac{N - x_{+i}}{\sum_p \frac{\widehat{\nu}_i \gamma_{x_{p+}}^{(i)}(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\widehat{\nu}_i \gamma_{x_{p+}}^{(i)}(\boldsymbol{\epsilon}, \boldsymbol{\nu}) + \widehat{\epsilon}_i \gamma_{x_{p+}-1}^{(i)}(\boldsymbol{\epsilon}, \boldsymbol{\nu})}} = \widehat{\nu}_i \frac{N - x_{+i}}{N - \mathcal{E}(X_{+i} | \mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})} \quad (28)$$

where we recognize the updates obtained before for the original and recoded data. Deriving from these equations the parameter update for the identifiable parameter ϵ_i/ν_i (assuming that $\epsilon_1/\nu_1 = 1$) we obtain

$$\begin{aligned} \frac{\epsilon_i}{\nu_i} &= \frac{\widehat{\epsilon}_i}{\widehat{\nu}_i} \frac{x_{+i}}{\mathcal{E}(X_{+i} | \mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})} \frac{N - \mathcal{E}(X_{+i} | \mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})}{N - x_{+i}} \\ &= \frac{\widehat{\epsilon}_i}{\widehat{\nu}_i} \frac{\frac{x_{+i}}{N - x_{+i}}}{\frac{\mathcal{E}(X_{+i} | \mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})}{N - \mathcal{E}(X_{+i} | \mathbf{x}_{.+}, \widehat{\boldsymbol{\epsilon}})}} \end{aligned} \quad (29)$$

We find that combining the two updates derived from the recoding of the observations yields yet another monotone algorithm for computing CML estimates. This algorithm is similar to the implicit equation algorithm except that the update factor in (29) is the ratio of observed and expected *odds-ratios*.

It is readily seen that the new algorithm in Equation 29 changes a parameter at least as much as the original implicit equation algorithm and both algorithms change the parameter in the same direction. Theorem 2 shows that also for the update in Equation 29, the new value is between the current value and the value at which the conditional log-likelihood attains its largest value. Hence, this new update leads to a higher likelihood value than the original implicit equation algorithm, and the variant of the implicit equation algorithm introduced before.

By the same reasoning we can also derive an improved version of the implicit equation algorithm for computing marginal maximum likelihood estimates. This algorithm also involves the observed and expected odds-ratio,

albeit with a different expectation. Note that the approach used to speed up the convergence rate (i.e., deliberately overparameterizing the model) is similar to the one used by Liu, Rubin, and Wu (1998) to accelerate the EM algorithm.

5 More complicated models

Essentially the same steps that led to the implicit equation algorithm for the Rasch model can be used to derive algorithms for ML estimation of considerably more general exponential family models. As an illustration, we consider a special case of the *Nominal Response Model* (Bock, 1972). Consider an item i with $m_i + 1$ response alternatives $j = 0, \dots, m_i$; one of which is chosen. Let X_{pi} denote the response alternative and for practical reasons we also consider the dummy coded variables $Y_{pij} = 1$ if category j was chosen and $Y_{pij} = 0$ otherwise. The item response function of the NRM is given by

$$\begin{aligned} P(Y_{pij} = y_{pij} | \theta_p) &= \frac{\exp [y_{pij}(a_{ij}\theta_p - \delta_{ij})]}{\sum_h \exp(a_{ih}\theta_p - \delta_{ih})} \\ &= \frac{\tau_p^{a_{ij}y_{pij}} \epsilon_{ij}^{y_{pij}}}{\sum_h \tau_p^{a_{ih}} \epsilon_{ih}} \quad , \end{aligned}$$

where $\epsilon_{ij} = \exp(-\delta_{ij})$, and $a_{i0} = \delta_{i0} = 0$ (or $\epsilon_{i0} = 1$) for identification. If the parameters a_{ij} are known, the NRM specializes to an exponential family model in which $y_{p++} = \sum_i \sum_j a_{ij} y_{pij} = \sum_i a_{i,x_{pi}}$ is a sufficient statistic for θ_p . If furthermore, the parameters a_{ij} are assumed to be integer valued, such that more response patterns correspond to the same value for the sufficient statistic, the method of conditional maximum likelihood estimation can be

used. Among others the *One Parameter Logistic Model* (OPLM: Verhelst & Glas, 1995) and the partial credit model (e.g., Masters, 1982; Andersen, 1995, p. 280) are special cases that satisfy these additional constraints.

The conditional loglikelihood can be written as

$$\sum_i \sum_j y_{+ij} \ln \epsilon_{ij} - \sum_s n_s \ln \gamma_s \quad , \quad (30)$$

where s indexes values of the sufficient statistic y_{p++} . In this case, the elementary symmetric functions satisfy the recursion

$$\gamma_{y_{p++}} = \gamma_{y_{p++}}^{(j)} + \sum_{h=1}^{m_i} \epsilon_{jh} \gamma_{y_{p++}-a_{jh}}^{(j)} \quad (31)$$

These formulae specialize to those for the dichotomous Rasch model when $m_i = 1$ and $a_{ij} = 1$ for $j > 1$.

A minorization algorithm is found in the same way as before, because the recursion in Equation 31 is a linear function in ϵ_{ij} . Now, the category parameters can be updated one at the time and the update is of the same form as in the Rasch model:

$$\epsilon_{ij} = \hat{\epsilon}_{ij} \frac{y_{+ij}}{\sum_s n_s \frac{\hat{\epsilon}_{ij} \hat{\gamma}_{s-a_{ij}}}{\hat{\gamma}_s^{(i)} + \sum_{h=1}^{m_i} \hat{\epsilon}_{ih} \hat{\gamma}_{s-a_{jh}}^{(i)}}} = \hat{\epsilon}_{ij} \frac{y_{+ij}}{\mathcal{E}[Y_{+ij} | \mathbf{y}_{.++}, \hat{\boldsymbol{\epsilon}}]}$$

In the psychometric literature this algorithm is described by Andersen (p. 277 and p. 280, 1995). Again, without a proof that the algorithm is monotonely convergent.

A variant based on overparameterization is obtained by updating ϵ_{i0} along with the others and identifying the model afterwards. Thus, we obtain:

$$\epsilon_{ij} = \hat{\epsilon}_{ij} \frac{y_{+ij} \mathcal{E}[Y_{+i0} | \mathbf{y}_{.++}, \hat{\boldsymbol{\epsilon}}]}{y_{+i0} \mathcal{E}[Y_{+ij} | \mathbf{y}_{.++}, \hat{\boldsymbol{\epsilon}}]} \quad (32)$$

for $j = 1, \dots, m_i$.

Note that, in contrast to the Rasch model, conditions for the existence of finite conditional maximum likelihood estimates have not been described for the NRM. This makes an interesting topic for future research.

6 Discussion

This paper has barely scratched the surface when it comes to the use of minorization based methods for constructing monotonely convergent algorithms for computing maximum likelihood, or Bayes modal, estimates. We have only considered methods based on simple concavity properties of parts of the log-likelihood function. The new algorithms considered in this paper involve unidimensional, and in one instance two-dimensional, linear minorizations. Both the unidimensionality and the linearity can be relaxed. Fruitful work remains to be done in both areas. Especially interesting is the use of quadratic minorizations (e.g., Groenen, Giaquinto, & Kiers, 2003; van Ruitenburg, 2005; de Leeuw, 2006).

Why bother with minorization methods such as the ones developed in this paper, and not just use the Newton-Raphson algorithm to mechanically find the maximum likelihood estimates? First, such methods are not foolproof, and require good starting values. This is all the more true with the sort of logistic models, such as the Rasch model, that are considered in this paper. The log-likelihood for such models becomes linear as the parameters tend to plus or minus infinity. This implies a vanishing second derivative, and divergence problems for the Newton-Raphson algorithm. Second, methods

like the Newton-Raphson algorithm require the computation of a full matrix of second derivatives, and the solution of a linear system of M equations in M unknowns. Both become computationally demanding and unstable as M becomes large. Large numbers of items are not uncommon (e.g., $M > 5000$) in educational measurement contexts.

For the problem at hand, the approach of overparameterizing the model and deriving a minorization algorithm for the overparameterized model has yielded promising results. We were able, in this way, to derive a new algorithm that, all other things being equal, leads to a bigger gain in likelihood, compared to Fischer's implicit equation algorithm. This does not imply, however, that all other things not being equal, we may not overshoot the maximum value of the likelihood from one iteration to the next. That is, if we update parameter ϵ_j , and next also update all the other parameters, we may have changed the sign of the derivative with respect to ϵ_j . Hence, we can not guarantee that the new algorithm *always* converges faster.

In closing, we remark that minorization is not the only principle from which to derive simple algorithms that are monotonely convergent. For example, van Ruitenburg (2005) obtained promising results with methods based on the iterative refinement of an interval that is constructed to contain the maximum likelihood estimate, such as the classical *false position* method.

Appendix

Theorem 2. *If $\bar{\epsilon}_j$ denotes the value at which the conditional log-likelihood for the Rasch model assumes its largest value, when considered as a function of ϵ_j , then*

$$\epsilon_j = \hat{\epsilon}_j \frac{\frac{x_{+j}}{N-x_{+j}}}{\frac{\mathcal{E}(X_{+j}|\mathbf{x}_{+},\hat{\epsilon})}{N-\mathcal{E}(X_{+j}|\mathbf{x}_{+},\hat{\epsilon})}}$$

is in-between $\hat{\epsilon}_j$ and $\bar{\epsilon}_j$.

Proof. We proof the theorem by showing that the sign of the derivative of the conditional log-likelihood:

$$\frac{\partial}{\partial \epsilon_j} l_{\mathbf{x}|\mathbf{x}_{+}}(\boldsymbol{\epsilon}) = \frac{1}{\epsilon_j} x_{+i} - \frac{1}{\epsilon_j} \mathcal{E}(X_{+j})$$

with respect to ϵ_j is the same for $\hat{\epsilon}_j$ and the new value. Without loss of generality, we assume that $\hat{\epsilon}_j < \bar{\epsilon}_j$, which implies that the sign of the partial derivative at $\hat{\epsilon}_j$ is positive. As a consequence, the new value will be larger than $\hat{\epsilon}_j$. From the monotonicity of the odds function, we immediately obtain that:

$$\text{sign} \left(\frac{\partial}{\partial \epsilon_j} l_{\mathbf{x}|\mathbf{x}_{+}}(\boldsymbol{\epsilon}) \right) = \text{sign} \left(\frac{1}{\epsilon_j} \frac{x_{+i}}{N-x_{+i}} - \frac{1}{\epsilon_j} \frac{\mathcal{E}(X_{+j})}{N-\mathcal{E}(X_{+j})} \right)$$

Filling in the new value in the first term of the right hand side equation yields that:

$$\text{sign} \left(\frac{1}{\epsilon_j} \frac{x_{+i}}{N-x_{+i}} - \frac{1}{\epsilon_j} \frac{\mathcal{E}(X_{+j})}{N-\mathcal{E}(X_{+j})} \right) = \text{sign} \left(\frac{1}{\hat{\epsilon}_j} \frac{\hat{\mathcal{E}}(X_{+j})}{N-\hat{\mathcal{E}}(X_{+j})} - \frac{1}{\epsilon_j} \frac{\mathcal{E}(X_{+j})}{N-\mathcal{E}(X_{+j})} \right)$$

From Lemma 3 we know that the function

$$\frac{1}{\epsilon_j} \frac{\mathcal{E}(X_{+j})}{N-\mathcal{E}(X_{+j})}$$

is monotonely decreasing, and hence we find that $\epsilon_j > \hat{\epsilon}_j$ makes the derivative positive, as was to be shown. \square

Lemma 3. *The function*

$$\frac{1}{\epsilon_j} \frac{\mathcal{E}(X_{+j}|\mathbf{x}_{.+}, \boldsymbol{\epsilon})}{N - \mathcal{E}(X_{+j}|\mathbf{x}_{.+}, \boldsymbol{\epsilon})}$$

is monotonely decreasing in ϵ_j .

Proof.

$$\begin{aligned} \frac{1}{\epsilon_j} \frac{\mathcal{E}(X_{+j}|\mathbf{x}_{.+}, \boldsymbol{\epsilon})}{N - \mathcal{E}(X_{+j}|\mathbf{x}_{.+}, \boldsymbol{\epsilon})} &= \frac{\sum_s n_s \frac{\gamma_{s-1}^{(j)}}{\gamma_s^{(j)} + \gamma_{s-1}^{(j)} \epsilon_j}}{\sum_s n_s \frac{\gamma_s^{(j)}}{\gamma_s^{(j)} + \gamma_{s-1}^{(j)} \epsilon_j}} \\ &= \sum_s \frac{\gamma_{s-1}^{(j)}}{\gamma_s^{(j)}} \frac{n_s \frac{\gamma_s^{(j)}}{\gamma_s^{(j)} + \gamma_{s-1}^{(j)} \epsilon_j}}{\sum_t n_t \frac{\gamma_t^{(j)}}{\gamma_t^{(j)} + \gamma_{t-1}^{(j)} \epsilon_j}} \end{aligned}$$

In the second part we recognize:

$$P(X_{.j} = 0 | X_{.+} = s, \epsilon_j) = \frac{\gamma_s^{(j)}}{\gamma_s^{(j)} + \gamma_{s-1}^{(j)} \epsilon_j}$$

and $P(X_{.+} = s) = \frac{n_s}{N}$ such that

$$P(X_{.j} = 0) = \sum_s \frac{n_s}{N} \frac{\gamma_s^{(j)}}{\gamma_s^{(j)} + \gamma_{s-1}^{(j)} \epsilon_j}$$

Using Bayes' Theorem we obtain that

$$P(X_{.+} = s | X_{.j} = 0; \epsilon_j) = \frac{n_s \frac{\gamma_s^{(j)}}{\gamma_s^{(j)} + \gamma_{s-1}^{(j)} \epsilon_j}}{\sum_t n_t \frac{\gamma_t^{(j)}}{\gamma_t^{(j)} + \gamma_{t-1}^{(j)} \epsilon_j}}$$

It follows that the function of interest is the expectation over the distribution of $X_{.+} | X_{.j} = 0$ of the ratio $\gamma_{s-1}^{(j)} (\gamma_s^{(j)})^{-1}$ which is known to be a monotone function of s (Verhelst et al., 1984).

Monotone likelihood ratio for $P(X_{.+} = s|X_{.j} = 0; \epsilon_j)$ in ϵ_j follows readily from

$$\frac{P(X_{.j} = 0|X_{.+} = s_2, \epsilon_j)}{P(X_{.j} = 0|X_{.+} = s_1, \epsilon_j)} = \frac{\gamma_{s_2}^{(j)} \gamma_{s_1}^{(j)} + \gamma_{s_1-1}^{(j)} \epsilon_j}{\gamma_{s_1}^{(j)} \gamma_{s_2}^{(j)} + \gamma_{s_2-1}^{(j)} \epsilon_j}$$

which is seen to be monotone in ϵ_j because $\gamma_{s-1}^{(j)}(\gamma_s^{(j)})^{-1}$ is a monotone function in s . The result follows from the symmetry of monotone likelihood ratio and the fact that monotone likelihood ratio implies stochastic ordering (e.g., Ross, 1996, chapter 9). \square

References

- Andersen, E. B. (1972). The numerical solution to a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, *34*, 283-301.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69-81.
- Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North-Holland Publishing Company.
- Andersen, E. B. (1995). Polytomous rasch models and their estimation. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. foundations, recent developments, and applications* (p. 271-291). New York: Springer-Verlag.
- Bock, R. D. (1972). Estimating item parameters and latent ability when

- responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- de Leeuw, J. (2006). *Some majorization techniques* (Tech. Rep. No. 2006032401). Department of Statistics, UCLA.
- De Leeuw, J. (1994). Block-relaxation algorithms in statistics. In W. L. H. H. Bock & M. M. Richter (Eds.), *Information systems and data analysis* (p. 308-325). Berlin: Springer-Verlag.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society*, 39, 1-38.
- Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Verlag Hans Huber. (Introduction to the theory of psychological tests.)
- Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, 46, 59-77.
- Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (p. 131-155). Berlin, Germany: Springer.
- Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, p. 515-585). Amsterdam, The Netherlands: Elsevier.

- Groenen, P. J. F., Giaquinto, P., & Kiers, H. A. L. (2003). *Weighted majorization algorithms for weighted least squares decomposition models* (Econometric Institute Report No. 9). Erasmus University Rotterdam.
- Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics*, *5*, 815-841.
- Hardy, G. H., Littlewood, J. E., & Polya, G. (1952). *Inequalities* (2 ed.). Cambridge: University Press.
- Heiser, W. J. (1995). Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. In W. J. Krzanowski (Ed.), *Recent advances in descriptive multivariate analysis* (p. 157-189). Oxford University Press.
- Lange, K., Hunter, D., & Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, *9*, 1-20.
- Liu, C., Rubin, D., & Wu, Y. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, *85*, 755-770.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.
- 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: The Danish Institute of Educational Research.
(Expanded edition, 1980. Chicago, The University of Chicago Press)
- Ross, S. M. (1996). *Stochastic processes* (Second ed.). New York: John Wiley & sons.
- van Ruitenburg, J. (2005). *Algorithms for parameter estimation in the Rasch model* (Tech. Rep. No. 05-04). Cito.
- Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (p. 215-238). New York: Springer Verlag.
- 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(3), 245-262.