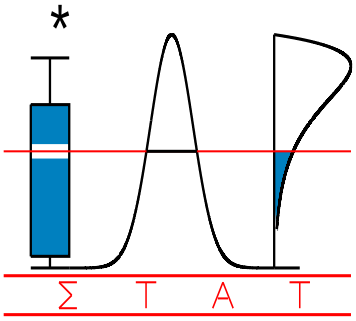
T E C H N I C A L
R E P O R T

0351

BAYESIAN ESTIMATION OF BASIC IRT MODELS

DE KNOP, S., RIJMEN, F. and I. VAN MECHELEN



I A P S T A T I S T I C S
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Bayesian estimation of basic IRT models

Stijn De Knop, Frank Rijmen and Iven Van Mechelen

Katholieke Universiteit Leuven

Author Notes:

The research reported in this paper was supported by grant IAP P5/24. Correspondence concerning this paper should be addressed to Iven Van Mechelen, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium. Email: Iven.VanMechelen@psy.kuleuven.ac.be.

Running head: Bayesian estimation of basic IRT models: an overview

Abstract

This paper presents an evaluation of different Bayesian estimation approaches for four basic IRT models: the Rasch model, the linear logistic test model, and their probit link counterparts. As to estimation algorithms, we focus on a data augmentation algorithm (with and without transformation of the parameters and the latent data) that includes a Gibbs sampling procedure, and on a Metropolis algorithm. A comparison between the algorithms for the different models is made on the level of convergence speed and recovery of the underlying truth. Differences in convergence speed appeared to be considerable; an explanation for this finding is provided.

1 Introduction

Over the last decade, interest in fully Bayesian estimation methods for psychometric models (and statistical models more in general), has widely increased (e.g., Gilks, Richardson, & Spiegelhalter (1996)). The reason for this is at least threefold: (1) Firstly, the Bayesian framework is very flexible with respect to the estimation of more elaborated models; estimation algorithms for basic models can easily be extended to account for model expansions, mostly without facing difficult analytical issues. (2) Secondly, the Bayesian approach provides a much broader view on the whole of the parameter space unlike standard frequentistic or maximum likelihood approaches; as such, a comprehensive view on estimates' uncertainty can be obtained without relying on asymptotic approximations. (3) Thirdly, one may note the advantages of the flexible inferential framework as implied by a Bayesian approach; by means of posterior predictive check procedures, for instance, one may easily check absolute and relative goodness of fit as well as specific model assumptions.

Different methods have been developed in order to estimate model parameters within a Bayesian framework. While some of these methods focus on the direct estimation of the posterior mode (a),

others allow for summaries of the posterior parameter distribution on the basis of simulated samples as obtained via MCMC methods (b).

(a) In the context of IRT models, with regard to mode seeking procedures Bock and Aitkin (1981) developed an EM algorithm that gives maximum likelihood estimates of item parameters of the marginal distribution as obtained by integrating over the ability distribution. Swaminathan and Gifford (1982, 1985, 1986) estimated the joint posterior modes for 1-, 2-, and 3-parameter logistic models by means of a Newton-Raphson algorithm. Mislevy (1986) also estimated the latter models, yet by means of an EM algorithm. Tsutakawa and Lin (1986) estimated the marginal posterior mode of the item parameters for the 2-parameter model. Tsutakawa and Soltys (1988) and Tsutakawa and Johnson (1990) provided analytical approximations for posterior means and variances of the person parameters, conditional on the item parameters, while the uncertainty of the latter is taken into account.

(b) As for sampling-based methods for Bayesian estimation of basic IRT models, one may refer to Albert (1992) and Albert and Chib (1993), who proposed a DA-Gibbs algorithm to simulate draws from the posterior distribution of a 2-parameter normal-ogive model. This methodology was extended by Verguts and De Boeck in order to estimate the probit counterpart of the linear logistic test model (LLTM) (Verguts & De Boeck, 2000). From their part, Maris and Maris (2002) showed how, for a Rasch model, one can sample from the posterior distribution by introducing a transformation of the parameters and the latent data in the DA-Gibbs sampler; the resulting algorithm is referred to as the DA-T-Gibbs sampler. Patz and Junker showed how a 2-parameter logistic model can be estimated using a Metropolis-Hastings sampling (Patz, 1996; Patz & Junker, 1999).

In the present paper we focus on fully Bayesian sampling-based estimation methods. In particular, we want to evaluate three different such methods in the context of basic IRT methods. The models under consideration are the Rasch model, the linear logistic test model and their probit-link

counterparts. The evaluation will be done both in terms of convergence speed and recovery of the underlying truth.

Introducing the notation used throughout the remainder of the paper the Rasch model can be summarized as

$$Pr\left(Y_{ni}=1|\theta_n,\beta_i\right)=logit^{-1}\left(\theta_n-\beta_i\right), \tag{1}$$

where θ_n is the ability of person n, with $1 \leq n \leq N$, β_i denotes the difficulty of item i, with $1 \leq i \leq I$, and Y_{ni} represents the response of person n to item i; it is further assumed that $p\left(\theta_n\right) \propto \mathcal{N}\left(0,\sigma_\theta^2\right)$. The probit counterpart of the Rasch model is

$$Pr\left(Y_{ni}=1|\theta_n,\beta_i\right)=\Phi\left(\theta_n-\beta_i\right). \tag{2}$$

The linear logistic test model is a Rasch model where the item parameters can be written as a linear combination of a restricted number of basic parameters η_j, with $j=1,\ldots,J$. The weights of the constituent basic parameters are assumed to be known a priori, and are represented in the $I \times J$ matrix $\mathbf{Q}$, such that $\beta_i=\mathbf{q}_i\boldsymbol{\eta}_j$.

The remainder of this paper is organized as follows: In Section 2 the different estimation algorithms are introduced. In Section 3 the design of our simultation study as well as its results are described, and an explanation for the differences in convergence speed between the proposed algorithms is given. The paper is concluded by a discussion, presented in Section 4.

2 Estimation algorithms

The Bayesian sampling-based methods to be discussed will make use of two different MCMC procedures: the Gibbs sampler and the Metropolis-Hastings algorithm. We will now successively discuss the methods based on each of those two procedures.

2.1 Algorithms based on the Gibbs sampler

Within this section, we will describe two estimation algorithms, based on the Gibbs sampler, in order to obtain the posterior distribution of the parameters of the LLTM and the Rasch models; the first algorithm (DA-Gibbs) will further be used to estimate the probit-link versions of the models, whereas the second will be used to estimate both the logit-link and the probit-link versions.

2.1.1 The DA-Gibbs sampler

We will describe here a DA-Gibbs algorithm for the probit-link LLTM, the algorithm for the probit-link Rasch model being fully similar. The prior distribution of the basic parameters is chosen to be non-informative, that is, locally uniform, $p(\eta_j) = U(-\infty, +\infty)$, the truncation at -5 and 5 being chosen to avoid impropriety of the posterior; furthermore, for the ability variance we assume $p\left(\sigma_\theta^2\right) \propto \left(\sigma_\theta^2\right)^{-1}$.

The introduction of latent data into the Gibbs sampler (Albert, 1992; Albert & Chib, 1993) is based on the assumption that every observed data point y_{ni} corresponds to an underlying, unobserved data point z_{ni}. For the probit-link LLTM, the latent variable Z_{ni} is normally distributed around $\theta_n - \mathbf{q}_i^{\cdot}\boldsymbol{\eta}$, with a standard deviation of 1: $Z_{ni} \sim \mathcal{N}(\theta_n - \mathbf{q}_i^{\cdot}\boldsymbol{\eta}, 1)$. The relation between the latent data and the observed data is specified as follows:

$$y_{ni} = 1 \quad \text{iff} \quad z_{ni} \geq 0;$$

$$y_{ni} = 0 \quad \text{iff} \quad z_{ni} < 0.$$

Due to this data augmentation, the drawing from the full conditionals comes down to drawing from normal distributions. Indeed, the full conditional distributions for $\mathbf{z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are as follows:

$$p(z_{ni}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}) \propto \begin{cases} \mathcal{N}(\theta_n - \mathbf{q}_i^{\cdot}\boldsymbol{\eta}, 1), & \text{truncated at the left of 0 iff } y_{ni} = 1, \\ \mathcal{N}(\theta_n - \mathbf{q}_i^{\cdot}\boldsymbol{\eta}, 1), & \text{truncated at the right of 0 iff } y_{ni} = 0, \end{cases} \tag{3}$$

$$p\left(\theta_n|\mathbf{y},\mathbf{z},\boldsymbol{\eta}\right) \propto \mathcal{N}\left(\frac{\left(\sum_{i=1}^{I} z_{ni} + \mathbf{q}_i^{\cdot}\boldsymbol{\eta}\right) + \frac{\mu_\theta}{\sigma_\theta^2}}{I + \frac{1}{\sigma_\theta^2}}, \frac{1}{I + \frac{1}{\sigma_\theta^2}}\right), \tag{4}$$

$$p\left(\eta_{j'}|\mathbf{y},\mathbf{z},\boldsymbol{\theta},\boldsymbol{\eta}_{(j')}\right) \propto \tag{5}$$
$$\mathcal{N}\left(\frac{\left(\sum_{n=1}^{N}\sum_{i=1}^{I} q_{j'i}\left[\theta_n - z_{ni} + \sum_{j\neq j'}^{J} q_{ji}\eta_j\right]\right)}{N\sum_{i=1}^{I} q_{j'i}^2 + \frac{1}{\sigma_\theta^2}}, \frac{1}{N\sum_{i=1}^{I} q_{j'i}^2 + \frac{1}{\sigma_\theta^2}}\right),$$

where $\boldsymbol{\xi}_{(d)}$ denotes the parameter vector $\boldsymbol{\xi}$ with parameter ξ_d excluded. The variance σ_θ^2 of the distribution of the person parameters is drawn from:

$$p\left(\sigma_\theta^2|\mathbf{y},\boldsymbol{\theta}\right) \propto inv\chi^2\left(N, \frac{1}{N}\sum_{n=1}^{N}\theta_n^2\right). \tag{6}$$

2.1.2 The DA-T-Gibbs sampler

The DA-T-Gibbs sampler, as introduced by Maris and Maris (2002), can be used to obtain a sample from the posterior distribution of the parameters of the probit-link as well as the original logit-link Rasch and LLTM models. For the logit-link models, the DA-Gibbs sampler does not involve simple full conditionals for the parameters. The latter problem can be tackled by means of a transformation of the variables. As such, the sampling from the posterior distribution $p\left(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y}\right)$ is based on a rewriting of the models under study as models involving latent data (DA), and to subsequently transform the variables in order to simplify the conditional distributions (T). In the remainder of this section we will discuss the DA-T-Gibbs sampler when used as an estimation algorithm for the parameters of the original, logit-link LLTM. We assume the same prior distributions as in the previous section.

In order to reformulate the logit-link LLTM as a latent-data model, latent data X_{ni} are introduced that have a logistic distribution with mean $(\theta_n - \mathbf{q}_i^{'}\boldsymbol{\eta})$ and variance $\frac{\pi^2}{3}$. The relation between the observed and the latent data is defined as:

$$y_{ni} = 1 \quad \text{iff} \quad x_{ni} \geq 0;$$

$$y_{ni} = 0 \quad \text{iff} \quad x_{ni} < 0.$$

It then follows that Y_{ni} is Bernoulli distributed with:

$$\begin{aligned} p\left(Y_{ni} = 1 | \theta_n, \eta_j\right) &= \int_0^{\infty} \frac{e^{x_{ni} - \left(\theta_n - \mathbf{q}_i^{'}\boldsymbol{\eta}\right)}}{\left(1 + e^{x_{ni} - \left(\theta_n - \mathbf{q}_i^{'}\boldsymbol{\eta}\right)}\right)^2} \quad dx_{ni} \\ &= \frac{e^{\left(\theta_n - \mathbf{q}_i^{'}\boldsymbol{\eta}\right)}}{1 + e^{\left(\theta_n - \mathbf{q}_i^{'}\boldsymbol{\eta}\right)}}, \end{aligned}$$

what corresponds to a LLTM. Given the latter formulation of the LLTM, the posterior $p\left(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathbf{y}\right)$ can be written as $\int p\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{x} | \mathbf{y}\right) dx$, in which $p\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{x} | \mathbf{y}\right)$ is proportional to

$$\left(\prod_n \prod_i I_{(0,\infty)}\left(x_{ni}\right)^{y_{ni}} I_{(-\infty,0)}\left(x_{ni}\right)^{1-y_{ni}} \frac{e^{x_{ni} - \left(\theta_n - \mathbf{q}_i^{'}\boldsymbol{\eta}\right)}}{\left(1 + e^{x_{ni} - \left(\theta_n - \mathbf{q}_i^{'}\boldsymbol{\eta}\right)}\right)^2} \right) p\left(\boldsymbol{\theta}, \boldsymbol{\eta}\right), \tag{7}$$

where $I_X(x)$ is an indicator function that equals 1 if $x \in X$, and 0 otherwise, and $p\left(\boldsymbol{\theta}, \boldsymbol{\eta}\right)$ is the joint prior distribution of the ability parameter and the basic parameter; as to the latter, we assume independence, such that $p\left(\boldsymbol{\theta}, \boldsymbol{\eta}\right) = p\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\eta}\right)$.

The difficulty in using a Gibbs sampler to sample from (7) arises from the fact that the parameters appear in the distribution of the latent data. As such, complex full conditionals arise. In analogy with Maris and Maris (2002), we propose the following transformation in order to remove the parameters from the distribution of the latent data:

$$z_{ni} = x_{ni} - \left(\theta_n - \mathbf{q}_i^{'}\boldsymbol{\eta}\right). \tag{8}$$

After this transformation, the joint posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{x}|\mathbf{y})$ is proportional to

$$\left(\prod_n \prod_i I_{(0,\infty)} \left(z_{ni} + \theta_n - \mathbf{q}_i' \boldsymbol{\eta} \right)^{y_{ni}} I_{(-\infty,0)} \left(z_{ni} + \theta_n - \mathbf{q}_i' \boldsymbol{\eta} \right)^{(1-y_{ni})} \frac{e^{z_{ni}}}{\left(1 + e^{z_{ni}}\right)^2} \right) p(\boldsymbol{\theta}, \boldsymbol{\eta}), \tag{9}$$

For this joint posterior a Gibbs sampler can be constructed that subsequently draws the latent data, the basic parameters, the person parameters and the person parameter variance from their conditional distributions. For the latent data we note:

$$\begin{aligned} p(z_{ni}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}) &\propto I_{(0,\infty)} \left(z_{ni} + \theta_n - \mathbf{q}_i' \boldsymbol{\eta} \right) \tfrac{e^{z_{ni}}}{(1+e^{z_{ni}})^2} \text{ iff } y_{ni} = 1, \\ p(z_{ni}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}) &\propto I_{(-\infty,0)} \left(z_{ni} + \theta_n - \mathbf{q}_i' \boldsymbol{\eta} \right) \tfrac{e^{z_{ni}}}{(1+e^{z_{ni}})^2} \text{ iff } y_{ni} = 0. \end{aligned} \tag{10}$$

The full conditional distribution of the ability parameter θ_n is proportional to

$$\left(\prod_{i=1}^{I} I_{(0,\infty)} \left(z_{ni} + \theta_n - \mathbf{q}_i' \boldsymbol{\eta} \right)^{y_{ni}} I_{(-\infty,0)} \left(z_{ni} + \theta_n - \mathbf{q}_i' \boldsymbol{\eta} \right)^{1-y_{ni}} \right) p(\theta_n). \tag{11}$$

If we define a lower bound l_θ and an upper bound u_θ as

$$\begin{aligned} l_\theta &= max_{i:y_{ni}=1} \left(\mathbf{q}_i' \boldsymbol{\eta} - z_{ni} \right), \\ b_\theta &= min_{i:y_{ni}=1} \left(\mathbf{q}_i' \boldsymbol{\eta} - z_{ni} \right), \end{aligned} \tag{12}$$

it can easily be shown that the full conditional $p(\theta_n|\mathbf{y}, \mathbf{z}, \eta)$ is nothing more than the ability's prior, truncated at these bounds:

$$p(\theta_n|\mathbf{y}, \mathbf{z}, \eta) \propto I_{l_\theta, u_\theta}(\theta_n) \, p(\theta_n). \tag{13}$$

The LLTM-basic parameters, η_j (for $j = 1, \ldots, J$) , are subsequently drawn from their full conditional distributions. For the full conditional of $\eta_{j'}$ we note:

$$p\left(\eta_{j'} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\eta}_{(j')} \right) \propto \tag{14}$$

$$\left(\prod_{n=1}^{N} \prod_{i=1}^{I} \left[I_{(0,\infty)} \left(z_{ni} + \theta_n - q_{j'i} \eta_{j'} - \sum_{j \neq j'} q_{ji} \eta_j \right) \right]^{y_{ni}} \left[I_{(-\infty,0)} \left(z_{ni} + \theta_n - q_{j'i} \eta_{j'} - \sum_{j \neq j'} q_{ji} \eta_j \right) \right]^{1-y_{ni}} \right) p(\eta_{j'}).$$

As such, this distribution is the prior, truncated at $l_{\eta_{j'}}$ and $u_{\eta_{j'}}$, with

$$\begin{aligned} l_{\eta_{j'}} &= max \left(l_{1\eta_{j'}}, l_{2\eta_{j'}} \right), \\ u_{\eta_{j'}} &= min \left(u_{1\eta_{j'}}, u_{2\eta_{j'}} \right), \end{aligned} \tag{15}$$

where

$$
\begin{aligned}
l_{1\eta_{j'}} &= max_{(ni:y_{ni}=0 \wedge q_{j'i}>0)} \frac{\theta_n + z_{ni} - \sum_{j \neq j'} q_{ji}\eta_j}{q_{j'i}}, \\
l_{2\eta_{j'}} &= max_{(ni:y_{ni}=1 \wedge q_{j'i}<0)} \frac{\theta_n + z_{ni} - \sum_{j \neq j'} q_{ji}\eta_j}{q_{j'i}}, \\
u_{1\eta_{j'}} &= min_{(ni:y_{ni}=0 \wedge q_{j'i}<0)} \frac{\theta_n + z_{ni} - \sum_{j \neq j'} q_{ji}\eta_j}{q_{j'i}}, \\
u_{2\eta_{j'}} &= min_{(ni:y_{ni}=1 \wedge q_{j'i}>0)} \frac{\theta_n + z_{ni} - \sum_{j \neq j'} q_{ji}\eta_j}{q_{j'i}},
\end{aligned}
\tag{16}
$$

Note that the cases where $q_{ij'} = 0$ are dropped since then the inequalities posed by the conditional distribution of $\eta_{j'}$ are always fulfilled.

2.2 Metropolis-Hastings based procedures

Given the prior distributions as in Section 2.2.1, and given a symmetric normal jumping kernel $J\left(\boldsymbol{\xi}^*|\boldsymbol{\xi}^{t-1}\right) = \mathcal{N}\left(\boldsymbol{\xi}^*|\boldsymbol{\xi}^{t-1}, c\Sigma\right) = J\left(\boldsymbol{\xi}^{t-1}|\boldsymbol{\xi}^*\right)$ with Σ being a known variance matrix and c a scaling constant, a Metropolis within Gibbs algorithm for the LLTM may proceed as follows at iteration t:

1. Draw $\boldsymbol{\theta}^t \propto p\left(\boldsymbol{\theta}|\boldsymbol{\eta}^{t-1}, \mathbf{y}\right)$:

   (a) Draw $\theta^* \propto \mathcal{N}\left(\boldsymbol{\theta}^{t-1}, c\sigma_\theta^2\right)$. In order to define an appropriate jumping rule, fine tuning is needed at the level of the scaling parameter c. We chose to adapt c for each data set such that a mean acceptance ratio of approximately 40 percent is obtained (Gelman, Carlin, Stern, & Rubin, 1995).

   (b) Accept $\theta^t = \theta^*$ with probability

$$
\begin{aligned}
r &= min\{\tfrac{p(\boldsymbol{\theta}^*, \boldsymbol{\eta}|\mathbf{y})}{p\left(\boldsymbol{\theta}^{t-1}, \boldsymbol{\eta}|\mathbf{y}\right)}, 1\}; \\
&= min\{\tfrac{p\left(\mathbf{y}|\boldsymbol{\eta}^t, \boldsymbol{\theta}^*\right)p(\boldsymbol{\theta}^*)}{p\left(\mathbf{y}|\boldsymbol{\eta}^t, \boldsymbol{\theta}^{t-1}\right)p\left(\boldsymbol{\theta}^{t-1}\right)}, 1\}; \\
&= min\left\{\prod_n \prod_i \left[\frac{\frac{exp\left[y_{ni}\left(\theta_n^* - \mathbf{q}_i'\boldsymbol{\eta}^{t-1}\right)\right]}{exp\left[\theta_n^* - \mathbf{q}_i'\boldsymbol{\eta}^{t-1}\right]}}{\frac{exp\left[y_{ni}\left(\theta_n^{t-1} - \mathbf{q}_i'\boldsymbol{\eta}^{t-1}\right)\right]}{exp\left[\theta_n^{t-1} - \mathbf{q}_i'\boldsymbol{\eta}^{t-1}\right]}}\right], 1\right\}.
\end{aligned}
\tag{17}
$$

Otherwise, $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$.

2. $\forall j \in \{1, 2, \ldots, J\}$, draw $\eta_j^t \propto p\left(\boldsymbol{\eta}|\boldsymbol{\theta}^t, \boldsymbol{\eta}_{(<j)}^t, \boldsymbol{\eta}_{(>j)}^{t-1}, \mathbf{y}\right)$. In order to simplify notation, we will represent the matrix $\left[\boldsymbol{\eta}_{(<j)}^t, \eta_j^*, \boldsymbol{\eta}_{(>j)}^t\right]^T$ as $\boldsymbol{\eta}_{(<j,j,>j)}^*$, and the matrix $\left[\boldsymbol{\eta}_{(<j)}^t, \eta_j^{t-1}, \boldsymbol{\eta}_{(>j)}^t\right]^T$ as $\boldsymbol{\eta}_{(<j,j,>j)}^{t-1}$:

   (a) Draw $\eta_j^* \propto \mathcal{N}\left(\eta_j^{t-1}, c^{'}\sigma_\eta^2\right)$, where $c^{'}$ is again an appropriate scaling constant.

   (b) Accept $\eta_j^t = \eta_j^*$ with probability

$$
\begin{aligned}
r &= min\{\frac{p\left(\boldsymbol{\theta}^t, \boldsymbol{\eta}_{(<j,j,>j)}^*|\mathbf{y}\right)}{p\left(\boldsymbol{\theta}^t, \boldsymbol{\eta}_{(<j,j,>j)}^{t-1}|\mathbf{y}\right)}, 1\}; \\
&= min\{\frac{p\left(\mathbf{y}|\boldsymbol{\theta}^t, \boldsymbol{\eta}_{(<j,j,>j)}^*\right)p\left(\boldsymbol{\eta}_{(<j,j,>j)}^*\right)}{p\left(\mathbf{y}|\boldsymbol{\theta}^t, \boldsymbol{\eta}_{(<j,j,>j)}^{t-1}\right)p\left(\boldsymbol{\eta}_{(<j,j,>j)}^{t-1}\right)}, 1\}; \\
&= min\left\{\prod_n \prod_i \left[\frac{\frac{exp\left[y_{ni}\left(\theta_n^t - \mathbf{q}_i^{'}\boldsymbol{\eta}_{(<t,t,>t)}^*\right)\right]}{exp\left[\theta_n^t - \mathbf{q}_i^{'}\boldsymbol{\eta}_{(<t,t,>t)}^*\right]}}{\frac{exp\left[y_{ni}\left(\theta_n^{t-1} - \mathbf{q}_i^{'}\boldsymbol{\eta}_{(<t,t,>t)}^{t-1}\right)\right]}{exp\left[\theta_n^{t-1} - \mathbf{q}_i^{'}\boldsymbol{\eta}_{(<t,t,>t)}^{t-1}\right]}}\right], 1\right\}.
\end{aligned}
\tag{18}
$$

   Otherwise, $\eta_j^t = \eta_j^{t-1}$.

3. Draw the variance of the distribution of the person parameter:

$$
p\left(\sigma_\theta^2|\mathbf{y}, \boldsymbol{\theta}\right) \propto inv\chi^2\left(N, \frac{1}{N}\sum_{n=1}^{N}\theta_n^2\right). \tag{19}
$$

3 A comparison

3.1 The experimental design

As noted above, we restrict ourselves to a comparison of the performance of the presented algorithms, when used for the estimation of four basic IRT models: the Rasch model, the LLTM, and their probit-link counterparts. The models using the probit link function are estimated by all three algorithms, whereas, for reasons mentioned above, the logit-link Rasch model and LLTM are not

estimated by means of the DA-Gibbs sampler. As such, all algorithm-model combinations in the study can be summarized schematically as given in Table 1.

Table 1: The algorithm-model combinations under study

	Model	
Algorithm	Rasch	LLTM
DA	Probit-link	Probit-link
DAT	Probit- & Logit-link	Probit- & Logit-link
Metropolis	Probit- & Logit-link	Probit- & Logit-link

Note that the estimation of the probit-link LLTM and Rasch model by means of the DA-T-Gibbs sampler involves a transformation that is "redundant" in that these models can also be well estimated without it. The latter algorithm-model combinations are included, however, in order to improve the orthogonality of the experimental design, as well as to help us to uncover some difficulties related to the DA-T-Gibbs procedure.

To evaluate the different estimation procedures, the algorithms are run for simulated data sets differing in the number of persons (300 vs. 1000), the number of items (15 vs. 40), the number of basic parameters (3 vs. 6), and the degree of correlation between the columns of the design matrix, which was drawn from a multivariate normal distribution with a constant between-column correlation (amounting to .2 vs. 8.). Data are further generated, both under the probit- and logit-link LLTM. All this implies a design with 32 data type cells. We restrict ourselves to 10 replications per cell; the latter can be justified given that the within-cell variation will appear to be small for each of the data type cells under study.

3.2 Convergence

For each method m independent chains have been run, convergence being monitored using Gelman and Rubin's $\hat{\mathrm{R}}$ (Gelman & Rubin, 1992), with:

$$\sqrt{\hat{\mathrm{R}}} = \sqrt{\left(\frac{n-1}{n} + \frac{(m+1)\,B}{mnW}\right)}, \tag{20}$$

where B is the between-chain variance, and W the average of the m within-chain variances. The latter measure is calculated for the second half of the chains. If it falls below 1.2, the chains are considered to have converged. We chose a cut-off of 60000 iterations (corresponding to several hours of calculation time) to categorize an algorithm-model combination as converged or non-converged. Table 2 shows the means (and standard deviations) of the number of iterations needed for the different algorithm-model combinations to reach convergence for all combinations that always converged. Note that, for each algorithm-model combination, the mean and standard deviation is not only taken over the different replications per cell, but also over the different datatype cells; given that the within-cell variations per cell are small, the reported standard deviations are a first indication of the sensitivity of the algorithm-model combination to differences in the number of persons, items, etc. (see further below). If for a given cell (algorithm-model combination) nonconvergence occurred for at least one replication, this is indicated in Table 2 as 'nc'; note that if nonconvergence occurred in a cell, nonconvergence always occurred for almost all replications in that cell.

To examine the effect of the different design factors in our study on speed of convergence, a fixed effects analysis of variance was run for every model with the design factors and algorithm (restricted to the applicable, converged cells) as independent variables, and with the logarithm of the number of iterations needed for convergence as a dependent variable. (The logarithm with basis 10 was taken in order to improve the 'normality' of the data.) Table 3 represents all effects

Table 2: Mean number (and standard deviation) of iterations to reach convergence for all algorithm-model combinations

	Model			
	Rasch		LLTM	
Algorithm	Probit	Logit	Probit	Logit
DA	231 (95)	na	215 (100)	na
DAT	3118 (2838)	1254 (889)	nc	nc
Metropolis	nc	nc	350 (598)	119 (126)

'na': not applicable; 'nc': no convergence within 60000 iterations.

with effect size ω^2 exceeding .05. Note that all presented effects are statistically significant at the 5% level as well.

The main results with respect to convergence are the following:

- As for the probit Rasch model, Table 3 shows that the speed of convergence considerably depends on the estimation algorithm, with the DA algorithm converging much faster than its DAT counterpart; the latter difference could already be seen in Table 2. The influence of the link function of the model under which the data are simulated is further such that the data sets generated using the probit link need more than twice the number of iterations to converge than their logit link counterparts. With respect to the number of items, we note that more iterations are needed to reach convergence when more items are considered.

- The logit Rasch model could only be estimated comfortably by means of the DAT algorithm. Main determinants of the speed of convergence of this algorithm-model combination, appear to be the number of items (positively correlated with the number of iterations needed to reach convergence) and the link function in the model under which the data are simulated

Table 3: Effect sizes $\omega^2 > .05$ for analysis of variance of speed of convergence

	Model			
	Rasch		LLTM	
Effect	Probit ω^2	Logit ω^2	Probit ω^2	Logit ω^2
algo	.833	na	-	na
dlink	.081	.135	.572	-
item	.052	.828	-	.117
basis	-	-	.050	.198
cor	-	-	.092	.530
algo*dlink	-	na	.051	na
algo*cor	-	na	.081	na
basis*cor	-	-	-	.098

'algo': the estimation algorithm (for the Probit Rasch model, DA is compared to DAT, whereas for the Probit LLTM, DA is compared to Metropolis); 'dlink': the link function used to simulate the data (logit vs. probit); 'item': the number of items (15 vs. 40); 'basis': the number of basic parameters (3 vs. 6); 'cor': the correlation structure of the design matrix (mean correlation $\sim$.2 vs. $\sim$.8); 'na': not applicable (the respective model is estimated by one algorithm only); '-': no practical significance at the 5% level.

(data sets generated using a probit link need more time to get estimated).

- With respect to the probit LLTM model, it appears that probit-link-data sets converge slower than logit-link-data sets, and that models with smaller numbers of basic parameters, and models with lower level of correlation between the columns of the design matrix converged faster. The significant interaction effects represented in Table 3 are disordinal, with the DA algorithm needing less iterations than the DAT algorithm for the elevated state of the second independent variable (probit-link data resp. lower design matrix correlation) in both cases.

- The logit LLTM could only be estimated comfortably by means of the Metropolis algorithm. The speed of convergence of this algorithm-model combination is mainly determined by the positive correlation between, on the one hand, the number of iterations to reach convergence and, on the other hand, the number of items, the number of basic parameters, and the level of correlation between the columns of the design matrix; moreover, there appears to be an

ordinal interaction between the two latter independent variables in that the difference in speed of convergence between data sets differing in number of basic parameters is higher for a higher degree of correlation between the columns of the design matrix.

3.3 Recovery of the underlying truth

Apart from speed of convergence, it is important that the algorithm under consideration succeeds in obtaining good parameter estimates. The recovery of the underlying truth was checked by means of both the correlation and the mean difference between the estimates and the true parameters. For all converged algorithm-model combinations where the link function is identical to the one used for the generation of the data, it appears that the correlation between the true item parameters and the estimated item parameters is over 99 percent, whereas the corresponding correlation for the person parameters lies around 93 percent. If the link function in the model differs from the link function used for the generation of the data, the correlation between the true item parameters and the estimated item parameters stays at 99 percent, whereas the correlation for the person parameters falls down by approximately 6 percent.

To examine the effect of the design factors on the recovery of the underlying truth, analyses of variance were run for each model, with as dependent variables the mean absolute difference between the estimates and the true parameters for the item and person parameters respectively. For each of the different models Table 4 shows the effects that account for at least 5% of the variance of the dependent variables. As was to be expected, Table 4 shows that for both the item and the person parameters the link function of the model under which the data have been generated is the most important determinant of the quality of the estimates. In order to grasp the impact of a misspecification of the link function in the model, Table 5 presents the mean absolute differences between the estimates and the true item parameters for the different model/data-link

Table 4: Effect sizes $\omega^2 > .05$ for analyses of variance of the recovery of the underlying truth, both for the item parameter and for the person parameter

		Model			
		Rasch		LLTM	
Parameters	Effect	Probit ω^2	Logit ω^2	Probit ω^2	Logit ω^2
item parameter	algo	-	na	.114	na
	dlink	.913	.993	.589	.940
person parameter	dlink	.724	.872	.748	.885
	item	.261	.074	.244	.086
	dlink*item	-	.062	-	-

'algo': the estimation algorithm (for the Probit Rasch model, DA is compared to DAT;for the Probit LLTM, DA is compared to Metropolis); 'dlink': the link function used to simulate the data (logit vs. probit); 'item': the number of items (15 vs. 40); 'na': not applicable (the respective model is estimated by one algorithm only); '-': no practical significance at the 5% level.

combinations under study. As for the interpretation of the absolute figures given in Table 5, note that the mean range of the true item parameters equals approximately 4, and that the mean range of the true person parameters equals approximately 5.

From the results shown in Table 4, we further learn that no sizeable differences show up between the quality of the DA-estimates and the quality of the DAT-estimates, whereas such differences do show up between the DA-estimates of the item parameters and the Metropolis-estimates of these parameters, the Metropolis estimates being more accurate than their DA counterparts.

4 Discussion

Combining the results given under Section 3, allows one to to choose an optimal estimation algorithm, for a given IRT model:

Table 5: Mean absolute differences between estimates and true parameter for item and person parameters and different model/data link combinations under study.

		Model			
		Rasch		LLTM	
Parameters	Data-link	Probit	Logit	Probit	Logit
item parameter	probit	.10	.57	.12	.67
	logit	.36	.13	.38	.04
person parameter	probit	.28	.58	.27	.62
	logit	.46	.36	.46	.36

- When considering a probit Rasch model, the DA algorithm should be preferred over the DAT algorithm, since its convergence is significantly faster, and the quality of the estimates is comparable.
- When considering a probit LLTM, the Metropolis algorithm should be preferred above the DA algorithm, as both algorithms converge equally fast, and the Metropolis algorithm yields better estimates.
- When considering a logit Rasch one is bound to estimate by means of the DA-T-Gibbs algorithm.
- When considering a logit LLTM one is bound to estimate by means of a Metropolis algorithm.

Apart from the pragmatic consequences of the results of Section 3, these findings also leave us with several theoretical questions, including:

1. Why does the DA-T-Gibbs sampler converge much slower than the DA-T-Gibbs sampler,

especially when used in combination with the LLTM?

2. Why does the Metropolis algorithm only converge within a reasonable number of iterations when used to estimate an LLTM?

3. Why does a misspecification of the link function has such an impact?

1. The most significant difference is in speed of convergence between the DA and the DA-T-Gibbs sampler, especially when used to estimate the LLTM. The main cause for the latter difference is to be found in the different structure of the full conditionals of both the DA and the DAT algorithms. As outlined under Section 2.1, the full conditionals in the DA-Gibbs sampler are normal distributions, whereas in the DA-T-Gibbs sampler the full conditionals are truncated prior distributions. Equations 12 and 16 give the bounds for the truncation of the priors of θ and η, respectively. For the ability parameter θ, for instance, Equation 12 shows how the normal prior distribution is truncated at the left at the value of the most difficult of all correctly answered items and at the right at the value of the most easy of all incorrectly answered items. It is clear that, as more items are introduced, the 'slice' from which to sample is likely to narrow. In the Rasch model, the same argument holds for the full conditional of the item parameter β. For the LLTM, the argument holds as well; yet, as can be seen from Equation 16, the determination of the upper and lower bounds is now based on all persons and items instead of on one of both only. As such, the 'slices' from which the basic parameters η in the LLTM are sampled are likely to be much smaller than the ones the item parameters β in the Rasch model are sampled from. A trivial consequence of the narrowing of the 'slices' to sample from is a raise in the autocorrelation over iterations of that parameter, and thus a slower convergence. To quantify the difference in autocorrelation between the different algorithm-model combinations, Table 6 gives the mean autocorrelation

of the item parameter for these different combinations.

Table 6: Overview of the mean autocorrelation of the item parameter for all algorithm-model combinations

	Model			
	Rasch		LLTM	
Simulation algorithm	Probit	Logit	Probit	Logit
DA	.704	na	.748	na
DAT	.982	.981	.999	.999
Metropolis	.999	.999	.843	.796

'na': not applicable, as the given model cannot be estimated by the respective algorithm.

2. The finding that the Metropolis algorithm converges very bad for the Rasch model, compared to the LLTM, stems from the raise of dimensionality of the parameter space: When the dimensionality of this space rises, a proposed jump is harder to be accepted. In order to keep the acceptance ratio constant, the variance of the jumping then distribution will have to be taken smaller (Johnson & Albert, 1999). Consequently, the autocorrelation of the parameters will rise, resulting in a slower convergence.

3. The impact of a misspecification of the link function, as shown in Table 5, is a straightforward matter: the data link transforms the parameterspace into probabilities (binary data, more exactly, that represent these probabilities), whereas the model link transforms these probabilities back to the parameter space; it is trivial that better recovery occurs when these transformations are each others inverse.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669-679.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*, 443-459.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457-511.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain monte carlo in practice.* London: Chapman & Hall.

Johnson, V., & Albert, H., J. (1999). *Ordinal data modeling.* New York: Springer-Verlag.

Maris, G., & Maris, E. (2002). A mcmc-method for models with continuous latent responses. *Psychometrika*, *67*, 335-350.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177-195.

Patz, J. (1996). *Mcmc for item response theory models with applications for naep.* Unpublished doctoral dissertation, Carnegie Mellon University.

Patz, R., & Junker, B. (1999). A straightforward approach to markov cain monte carlo methods for item respons models. *Journal of Educational & Behavioral Statistics*, *24*, 146-178.

Swaminathan, H., & Gifford, J. (1982). Bayesian estimation in the rasch model. *Journal of Educational Statistics*, *9*, 175-191.

Swaminathan, H., & Gifford, J. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349-364.

Swaminathan, H., & Gifford, J. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, *51*, 589-601.

Tsustakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, *51*, 251-267.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, *55*, 371-390.

Tsutakawa, R. K., & Soltys, M. J. (1988). Approximation for bayesian ability estimation. *Journal of Educational Statistics*, *13*, 117-130.

Verguts, T., & De Boeck, P. (2000). A rasch model for detecting learning while solving an inteeligence test. *Applied Psychological Measurement*, *24*, 151-162.