Psicológica (2005), 26, 327-352

An Introduction to the DA-T Gibbs Sampler for the Two-Parameter Logistic (2PL) Model and Beyond

Gunter Maris & Timo M. Bechger
Cito (The Netherlands)

The DA-T Gibbs sampler is proposed by Maris and Maris (2002) as a Bayesian estimation method for a wide variety of *Item Response Theory (IRT) models*. The present paper provides an expository account of the DA-T Gibbs sampler for the 2PL model. However, the scope is not limited to the 2PL model. It is demonstrated how the DA-T Gibbs sampler for the 2PL may be used to build, quite easily, Gibbs samplers for other IRT models. Furthermore, the paper contains a novel, intuitive derivation of the Gibbs sampler and could be read for a graduate course on sampling.

Introduction

Let $Y_{pi} = 1$ denote the event that person p gives the correct answer to item i, and θ_p his ability. Assume that there exists a *latent response variable*, X_{pi}, such that person p solves item i if X_{pi} is larger than a threshold δ_i. That is,

$$P(Y_{pi} = 1|\theta_p) = P(X_{pi} > \delta_i|\theta_p) \quad .$$

It is seen that the probability of a correct response depends on the threshold of the item as well as the ability of the respondent. The probability of a correct response as a function of ability is called the *Item Response Function (IRF)*.

Address correspondence to: Gunter Maris, Cito, P.O. Box 1034, NL-6801 MG, Arnhem, The Netherlands. E-mail: gunter.maris[at]citogroep.nl; Tel:+31-026-3521162.

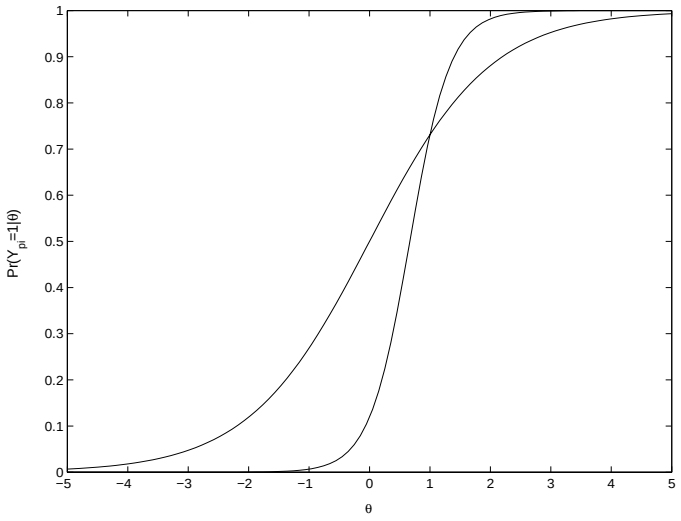

Figure 1. IRFs for two 2PL items with different parameters.

Under the *Two-Parameter Logistic (2PL) model* (Birnbaum, 1968), X_{pi} is assumed to follow a logistic distribution with mean $\alpha_i\theta_p$ and scale parameter $\beta = 1$ so that

$$\begin{aligned} P(X_{pi} > \delta_i|\theta_p, \alpha_i, \delta_i) &= \int_{-\infty}^{\infty} (x_{pi} > \delta_i) f(x_{pi}|\theta_p, \alpha_i) dx_{pi} \\ &= \int_{-\infty}^{\infty} (x_{pi} > \delta_i) \frac{\exp(x_{pi} - \alpha_i\theta_p)}{[1 + \exp(x_{pi} - \alpha_i\theta_p)]^2} dx_{pi} \\ &= \frac{\exp(\alpha_i\theta_p - \delta_i)}{1 + \exp(\alpha_i\theta_p - \delta_i)} \quad , \end{aligned}$$

where $(x_{pi} > \delta_i)$ denotes an indicator variable that is one if $x_{pi} > \delta_i$, and zero otherwise. The *Two-parameter Normal Ogive (2NO) model* (Birnbaum, 1968) is obtained when the distribution of the latent response variables is normal.

The *discrimination parameter* α_i determines how fast the IRF changes with ability. If α_i is positive (negative), the probability of answering correctly is an increasing (decreasing) function of ability. The *Rasch model* (Rasch, 1980) is a special case of the 2PL where all items have a discrimination parameter equal to one.

As it stands, the 2PL is *unidentifiable*. Specifically,

$$P(Y_{pi} = 1|\theta_p, \alpha_i, \delta_i) = \frac{\exp(\alpha_i^*\theta_p^* - \delta_i^*)}{1 + \exp(\alpha_i^*\theta_p^* - \delta_i^*)}$$

where

$$\alpha_i^* = \alpha_i d, \quad \delta_i^* = \delta_i - \alpha_i c, \quad \theta_p^* = \frac{\theta_p - c}{d} \quad ,$$

and c and d are arbitrary constants. To deal with this indeterminacy we arbitrarily set $\alpha_1 = 1$, and $\delta_1 = 0$. This means that the item parameters must be interpreted relative to the first item.

The purpose of this paper is to provide an expository account of Bayesian estimation of the 2PL focussing on the DA-T Gibbs sampler developed by Maris and Maris (2002). In addition, we offer an intuitive derivation of the Gibbs sampler and demonstrate that the DA-T Gibbs sampler for the 2PL can be used to build Gibbs samplers for many other *Item Response Theory (IRT)* models. Among others, we consider *the Linear Logistic Test Model (LLTM)* (Fischer, 1995), the 3PL (Birnbaum, 1968), and *the Nedelsky model* for multiple choice items (Bechger, Maris, Verstralen & Verhelst, 2005).

Gibbs Sampling

Let $\Lambda = (\Lambda_1, \ldots, \Lambda_m)$, $m \geq 2$, denote a vector of parameters.[1] In Bayesian statistics, the unknown parameters are considered random variables. Bayes theorem states that the *posterior density* (the posterior, for short) of Λ given the observed data $\mathbf{y}$ is given by

$$f(\lambda|\mathbf{y}) = \frac{f(\mathbf{y}|\lambda)f(\lambda)}{f(\mathbf{y})} \quad ,$$

where $f(\mathbf{y}|\lambda)$ denotes the likelihood function, and $f(\mathbf{y})$ the marginal likelihood function. The *prior density* $f(\lambda)$ (prior, for short) expresses substantive knowledge concerning the parameters prior to data collection. In Bayesian statistics all inferences about the parameters are based upon the posterior.

[1] We use subscripts to distinguish parameter vectors from scalars.

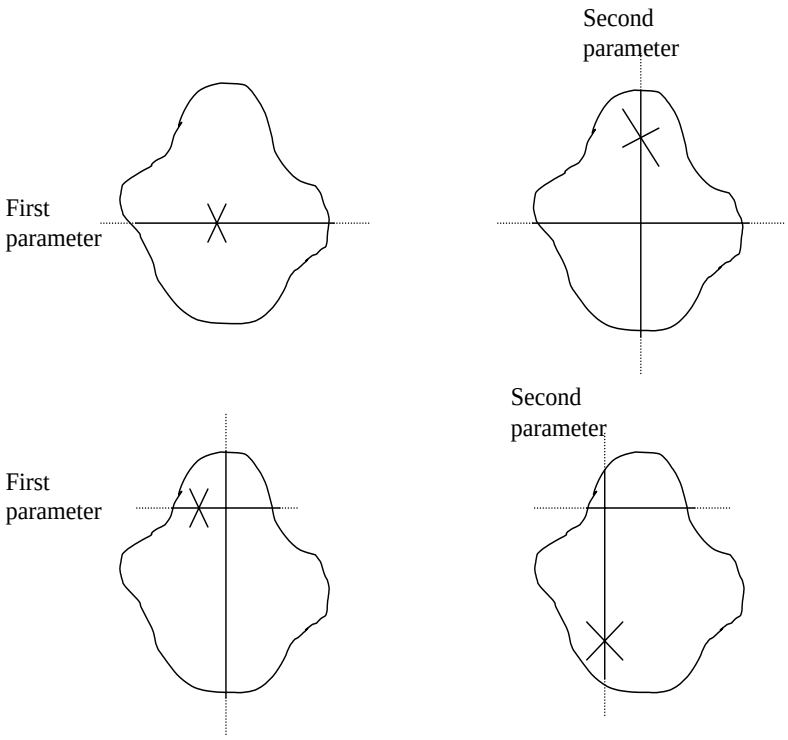

Figure 2. Schematic representation of two iterations of the Gibbs sampler with two parameters. The plot must be read from upper left to lower right.

The Gibbs sampler is an iterative procedure to generate parameter values $\lambda^{(0)}, \lambda^{(1)}, \ldots$ from the posterior. The first n generated values are discarded and the rest is considered to be a *dependent and identically distributed (did)* sample from the posterior. This means that

1. The distribution of $\Lambda^{(n+j)}$ given the data is the posterior for all $j > 0$.
2. Conditional upon the data, $\Lambda^{(n+j)}$ is *not* independent of $\Lambda^{(n+i)}$.

In this section we discuss *how* the Gibbs sampler works, *why* it works, and *when* it works. Alternative explanations can be found, for instance, in Casella and George (1992), Tanner (1996), or Ross (2003). The reader is referred to Tierney (1994) for a more rigorous treatment.

How

The procedure starts by choosing an initial value $\lambda^{(0)}$. Then, in each successive iteration, individual parameters are sampled independently from their so-called full conditional distributions. The order in which the parameters are sampled is arbitrary.

The *full conditional density* (the full conditional, for short) is the density given the observed data and the current value of all other parameters. In the sequel, we will often use the shorthand notation $f(\lambda_k|\dots)$ for the full conditional of parameter λ_k. To determine, up to a constant, the full conditional of a parameter λ_k we write down the density $f(\lambda, \mathbf{y})$ and remove all factors that are unrelated to λ_k. Specific examples will be given below.

Figure 2 represents two iterations of a fictitious Gibbs sampler, with two parameters being sampled at each iteration. The closed curve represent the *support* of a two-dimensional posterior. The solid lines indicate the support of the full conditionals and the crosses denote arbitrary values simulated from the different full conditional distributions. Observe that the Gibbs sampler "travels" through the support of the posterior along horizontal and vertical paths. Note that the support of the posterior must be such that every region can be reached by the Gibbs sampler, irrespective of the point of departure.

With a *did* sample from the posterior we may use the *Monte Carlo* method to calculate an unbiased estimate of the posterior expectation of any function $g(\lambda, \mathbf{y})$:

$$\int g(\lambda, \mathbf{y}) f(\lambda|\mathbf{y}) d\lambda \approx \frac{1}{n_s} \sum_j g(\lambda^{(j)}, \mathbf{y}) \quad ,$$

where n_s denotes the number of sampled values $\lambda^{(j)}$. That is, we approximate the expectation by the sample mean. The posterior probability that a parameter is smaller or equal to a constant t, for example, is estimated by

$$\frac{1}{n_s} \sum_j \left(\lambda_k^{(j)} \leq t\right) \quad .$$

The variance of the estimator of the posterior expectation can be estimated by the variance over independent replications of the Gibbs sampler.

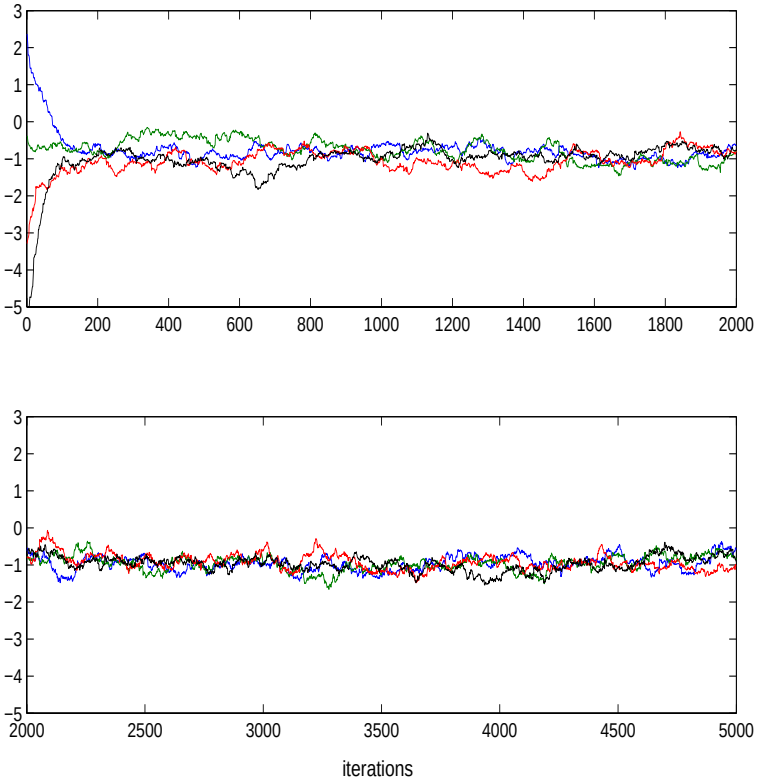

Figure 3. Plot of sampled values against iterations.

Unfortunately, there is no established way to determine an appropriate value for n. One option is to look at plots of $\lambda^{(1)}, \lambda^{(2)}, \ldots$ against iterations for a number of independent replications. An illustration with four independent replications is given in Figure 3. If, after n iterations, the values appear to fluctuate around a common stationary value, this may be taken as circumstantial evidence that n is large enough. In Figure 3, the plots appear to stabilize after about 1200 iterations. However, there is no way to be sure since we do not know what will happen after 5000 iterations. Other *ad hoc* methods to assess the required number of iterations are surveyed by Gelman and Rubin (1992) or Gill (2002, chapter 11).

Why

Let $\left\{\Lambda^{(n)}, n \geq 0\right\}$ denote a Markov chain. A *Markov chain* is a stochastic process such that each value depends only on its immediate predecessor; that is, for $n > 0$,

$$f(\lambda^{(n)}|\lambda^{(n-1)}, \ldots, \lambda^{(0)}) = f(\lambda^{(n)}|\lambda^{(n-1)}) \quad .$$

The Gibbs sampler is a procedure to simulate a Markov chain such that the marginal distribution of $\Lambda^{(n)}$ converges to the posterior if n increases. Convergence to the posterior is guaranteed if the following conditions are satisfied:

1. The posterior is the invariant distribution.
2. The chain is irreducible.

Invariance means that if $\lambda^{(0)}$ is drawn from the posterior, then all subsequent values are also draws from the posterior. Suppose, for ease of presentation, that there are two parameters.[2] Their posterior density is

$$f(\lambda_1, \lambda_2|\mathbf{y}) = f(\lambda_1|\lambda_2, \mathbf{y}) f(\lambda_2|\mathbf{y}) \quad .$$

To sample from this posterior, we draw $\lambda_2^{(1)}$ from the marginal posterior distribution and then $\lambda_1^{(1)}$ from the distribution conditional upon $\lambda_2^{(1)}$. Note that the latter is a full conditional as defined in the previous paragraph.

Suppose we set up a Markov chain to draw $\lambda_2^{(1)}$ from the marginal posterior distribution. Convergence is faster if the dependence between subsequent values is weaker. Thus we aim for a weak degree of dependence. Specifically, we ensure that $\Lambda_2^{(1)}$ and $\Lambda_2^{(0)}$ are independent and identically distributed conditional upon $\Lambda_1^{(0)}$. That is,

$$f(\lambda_2^{(0)}, \lambda_2^{(1)}|\mathbf{y}) = \int f(\lambda_2^{(0)}|\lambda_1^{(0)}, \mathbf{y}) f(\lambda_2^{(1)}|\lambda_1^{(0)}, \mathbf{y}) f(\lambda_1^{(0)}|\mathbf{y}) d\lambda_1^{(0)} \quad .$$

If we integrate $f(\lambda_2^{(0)}, \lambda_2^{(1)}|\mathbf{y})$ with respect to $\lambda_2^{(0)}$ (or $\lambda_2^{(1)}$), we see that $\Lambda_2^{(0)}$ and $\Lambda_2^{(1)}$ have the same marginal distribution. This distribution is the marginal

[2] The argument for the general case follows by mathematical induction.

posterior. It follows that

$$\begin{aligned} f(\lambda_2^{(1)}|\lambda_2^{(0)},\mathbf{y}) &= \frac{f(\lambda_2^{(0)},\lambda_2^{(1)}|\mathbf{y})}{f(\lambda_2^{(0)}|\mathbf{y})} \\ &= \int f(\lambda_2^{(1)}|\lambda_1^{(0)},\mathbf{y})\frac{f(\lambda_2^{(0)}|\lambda_1^{(0)},\mathbf{y})f(\lambda_1^{(0)}|\mathbf{y})}{f(\lambda_2^{(0)}|\mathbf{y})}d\lambda_1^{(0)} \\ &= \int f(\lambda_2^{(1)}|\lambda_1^{(0)},\mathbf{y})f(\lambda_1^{(0)}|\lambda_2^{(0)},\mathbf{y})d\lambda_1^{(0)} \quad . \end{aligned}$$

To produce a value $\lambda_2^{(1)}$ from the posterior distribution we may use the *method of composition* (Tanner, 1996, section 3.3.2) as follows:

1. Draw $\lambda_2^{(0)}$ from the posterior.
2. Draw $\lambda_1^{(0)}$ from the full conditional $f(\lambda_1|\lambda_2^{(0)},\mathbf{y})$.
3. Draw $\lambda_2^{(1)}$ from the full conditional $f(\lambda_2|\lambda_1^{(0)},\mathbf{y})$.

This procedure is a Gibbs sampler starting with a draw from the posterior.

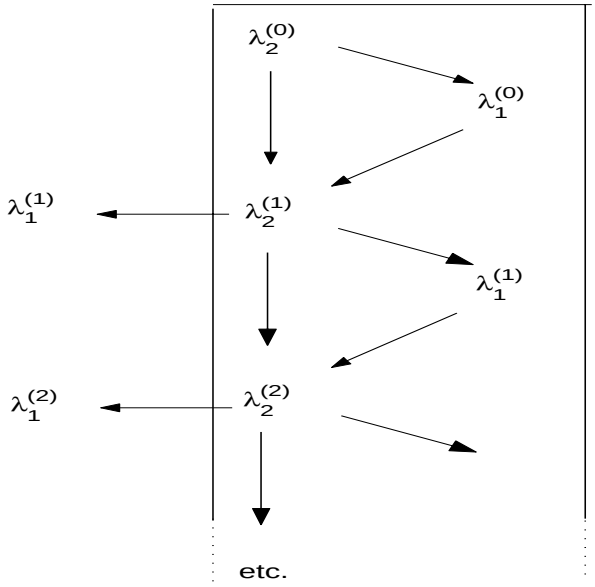

Figure 4. Schematic picture of the sampling procedure. Within the rectangle is the Gibbs sampler.

With $\lambda_2^{(1)}$ drawn from the marginal posterior, we then draw $\lambda_1^{(0)}$ from the full conditional $f(\lambda_1|\lambda_2^{(1)}, \mathbf{y})$ and repeat the process with $\lambda_2^{(1)}$ replacing $\lambda_2^{(0)}$, etc. Schematically, the sampling procedure may be depicted as in Figure 4 where the values generated by the Gibbs sampler are drawn inside a rectangle. It can be shown that these values are the realization of a Markov chain whose invariant distribution is, by construction, the posterior. The values outside the rectangle need not be generated in which case we obtain the Gibbs sampler as described in the previous section.

Irreducibility refers to the fact that it must be possible to reach each region in the support of the posterior (see e.g., Figure 2). This is true for the majority of applications.

When

Gibbs sampling is useful when the full conditionals are *tractable*. If so, it provides an estimation procedure that can be implemented relatively quickly. We call a distribution *tractable* if there is a simple and efficient method to generate a sample from it. Methods for stochastic simulation can be found, for instance, in Devroye (1986), Ripley (1997), or Ross (2001).

There are many situations where the full conditionals are not tractable. This is, in fact, the case of the 2PL (Maris & Maris, 2002). In the next section, we will demonstrate that the DA-T Gibbs sampler is a variant of the Gibbs sampler designed to give tractable full conditionals.

The DA-T Gibbs Sampler for the 2PL

The Prior

We conveniently assume that the parameters are *a priori* independent. That is,

$$f(\theta, \delta, \alpha) = \prod_p f(\theta_p) \prod_i f(\delta_i) f(\alpha_i) \quad . \tag{1}$$

We also assume that all prior distributions are tractable. Note that the priors must be chosen relative to the item whose parameters are arbitrarily fixed to identify the model.

The Full Conditionals

DA stands for *Data Augmentation* which entails adding latent data as (auxiliary) parameters (Tanner & Wong, 1987). The principle of DA may be stated as follows:

> Augment the observed data with latent data so that the augmented posterior distribution is "simple". (e.g., Tanner, 1996, p. 38)

Here, the continuous latent responses are added as parameters. Our hope is that DA will result in tractable full conditionals. Let's see !

The DA posterior of the 2PL is proportional to

$$\begin{aligned} f(\theta,\delta,\alpha,\mathbf{x},\mathbf{y}) &= f(\mathbf{y}|\mathbf{x},\theta,\delta,\alpha)f(\mathbf{x}|\theta,\delta,\alpha)f(\theta,\delta,\alpha) \\ &= f(\mathbf{y}|\mathbf{x},\delta)f(\mathbf{x}|\theta,\alpha)f(\theta,\delta,\alpha) \quad . \end{aligned}$$

Persons are assumed to be independent of one another, so that

$$\begin{aligned} f(\mathbf{y}|\mathbf{x},\delta) &= \prod_p \prod_i f(y_{pi}|x_{pi},\delta_i) \left(= \begin{cases} 1 \text{ if } x_{pi} > \delta_i \text{ and } y_{pi} = 1 \\ 1 \text{ if } x_{pi} \leq \delta_i \text{ and } y_{pi} = 0 \\ 0 \text{ otherwise} \end{cases} \right) \\ &= \prod_p \prod_i (x_{pi} > \delta_i)^{y_{pi}} (x_{pi} \leq \delta_i)^{1-y_{pi}} \quad , \end{aligned}$$

and

$$\begin{aligned} f(\mathbf{x}|\theta,\alpha) &= \prod_p \prod_i f(x_{pi}|\theta_p,\alpha_i) \\ &= \prod_p \prod_i \frac{\exp(x_{pi}-\alpha_i\theta_p)}{[1+\exp(x_{pi}-\alpha_i\theta_p)]^2} \quad . \end{aligned}$$

Thus $f(\theta,\delta,\alpha,\mathbf{x},\mathbf{y})$ equals

$$\prod_p \prod_i (x_{pi} > \delta_i)^{y_{pi}} (x_{pi} \leq \delta_i)^{1-y_{pi}} \frac{\exp(x_{pi}-\alpha_i\theta_p)}{[1+\exp(x_{pi}-\alpha_i\theta_p)]^2} f(\theta,\delta,\alpha) \quad . \quad (2)$$

The next step is to derive the full conditionals, including those for the latent responses. Let us first consider the full conditional distribution of δ_i,

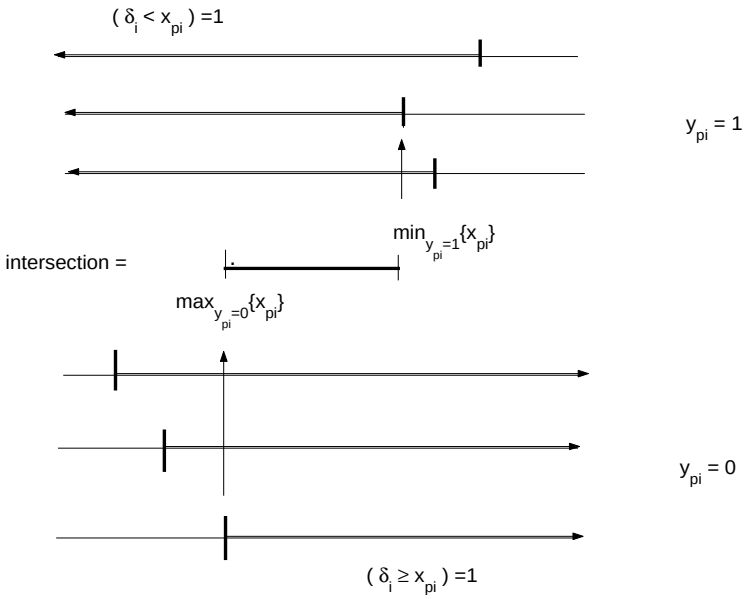

Figure 5. Illustration of the steps taken to arrive at Equation 4. It is illustrated that each factor in (3) represents a half open interval, extending either to $+\infty$ or $-\infty$, and their intersection is a closed interval.

where $i > 1$ since the first item location parameter was fixed. If we remove from (2) all terms that are unrelated to δ_i, we find that

$$f(\delta_i|\dots) \propto f(\delta_i)\left[\prod_p (x_{pi} > \delta_i)^{y_{pi}}(x_{pi} \leq \delta_i)^{1-y_{pi}}\right] \quad . \tag{3}$$

In Equation 3, the term within brackets represents a closed interval. It is illustrated in Figure 5 that

$$\begin{aligned}\prod_p (x_{pi} > \delta_i)^{y_{pi}}(x_{pi} \leq \delta_i)^{1-y_{pi}} &= \prod_{p:y_{pi}=1}(\delta_i < x_{pi}) \prod_{p:y_{pi}=0}(x_{pi} \leq \delta_i) \\ &= \left(\delta_i < \min_{p:y_{pi}=1}\{x_{pi}\}\right)\left(\max_{p:y_{pi}=0}\{x_{pi}\} \leq \delta_i\right) \\ &= \left(\max_{p:y_{pi}=0}\{x_{pi}\} \leq \delta_i < \min_{p:y_{pi}=1}\{x_{pi}\}\right) \quad . \end{aligned} \tag{4}$$

Thus, $f(\delta_i|\dots)$ is the truncated prior of δ_i which is tractable.

In a similar way, we find that the full conditionals of α_i and θ_p are proportional to the product of logistic with prior densities. For instance, the full conditional of any θ_p is

$$f(\theta_p|\dots) \propto f(\theta_p) \prod_i \frac{\exp(x_{pi} - \alpha_i\theta_p)}{[1+\exp(x_{pi} - \alpha_i\theta_p)]^2} \quad . \tag{5}$$

Unfortunately, the product of logistic densities is not tractable. We conclude that DA has not simplify the task of sampling from the full conditional distributions.[3]

As seen in Equation 5, the problem is due to the fact that the distribution of the latent responses depends on the item and person parameters. The DA-T Gibbs sampler is obtained if we transform the continuous latent responses to remove all parameters from their distribution. Hence, the T stands for *Transformation*. For the 2PL we apply the transformation $z_{pi} = x_{pi} - \alpha_i\theta_p$. The resulting "DA-T posterior", $f(\theta, \delta, \alpha, \mathbf{z}|\mathbf{y})$, is proportional to $f(\theta, \delta, \alpha, \mathbf{z}, \mathbf{y})$. From (2) it is easily found that $f(\theta, \delta, \alpha, \mathbf{z}, \mathbf{y})$ equals

$$\prod_p \prod_i (z_{pi} + \alpha_i\theta_p > \delta_i)^{y_{pi}} (z_{pi} + \alpha_i\theta_p \leq \delta_i)^{1-y_{pi}} \frac{\exp(z_{pi})}{[1+\exp(z_{pi})]^2} f(\theta, \delta, \alpha) \tag{6}$$

Removing unrelated factors from Equation 6 shows that each of the full conditionals is now a tractable truncated distribution:

1. The full conditional of z_{pi} is a logistic distribution with support

$$(z_{pi} > \delta_i - \alpha_i\theta_p)^{y_{pi}} \, (z_{pi} \leq \delta_i - \alpha_i\theta_p)^{1-y_{pi}}$$

2. The full conditional of δ_i ($i > 1$):

$$f(\delta_i|\dots) \propto f(\delta_i) \left[\prod_p (\delta_i < z_{pi} + \alpha_i\theta_p)^{y_{pi}} \, (\delta_i \geq z_{pi} + \alpha_i\theta_p)^{1-y_{pi}} \right]$$

3. The full conditional of α_i ($i > 1$):

$$f(\alpha_i|\dots) \propto f(\alpha_i) \left[\prod_p (\alpha_i\theta_p > \delta_i - z_{pi})^{y_{pi}} \, (\alpha_i\theta_p \leq \delta_i - z_{pi})^{1-y_{pi}} \right]$$

[3] In contrast, a product of normal densities is again a normal density. Thus, DA is effective for the 2NO model *with normal priors* (Albert, 1992; Albert & Chib, 1993). Here, we will not consider a method that works only for a particular prior distribution.

4. The full conditional of θ_p:

$$f(\theta_p|\dots) \propto f(\theta_p)\left[\prod_i \left(\theta_p\alpha_i > \delta_i - z_{pi}\right)^{y_{pi}}\left(\theta_p\alpha_i \leq \delta_i - z_{pi}\right)^{1-y_{pi}}\right]$$

In the following paragraph we demonstrate how the support of each full conditional is determined.

Calculating the Truncation Constants

The support of each of the full conditionals is seen to be a product of indicator functions of the following form:

$$\prod_j \left(l_j < \lambda_i \leq h_j\right) = \left(\max_j\left\{l_j\right\} < \lambda_i \leq \min_j\left\{h_j\right\}\right) \quad , \tag{7}$$

where either $l_j = -\infty$ or $h_j = \infty$. Hence, each term $(l_j < \lambda_i < h_j)$ restricts the range of λ_i to a half open interval extending to either plus or minus infinity. As illustrated in the previous section, their product is the intersection of these intervals, ranging from $\max_j\{l_j\}$ to $\min_j\{h_j\}$ (see Figure 5). Thus, $\max_j\{l_j\}$ and $\min_j\{h_j\}$ are the truncation constants for the full conditional.

The support for δ_i: The support for δ_i is a product of indicator functions over persons. We see that

$$l_p = \begin{cases} -\infty & \text{if } y_{pi} = 1 \\ z_{pi} + \alpha_i\theta_p & \text{if } y_{pi} = 0 \end{cases}$$

and

$$h_p = \begin{cases} z_{pi} + \alpha_i\theta_p & \text{if } y_{pi} = 1 \\ \infty & \text{if } y_{pi} = 0 \end{cases}$$

The support of α_i: Note that

$$\begin{aligned} &\left(\alpha_i\theta_p > \delta_i - z_{pi}\right)^{y_{pi}}\left(\alpha_i\theta_p \leq \delta_i - z_{pi}\right)^{1-y_{pi}} \\ &= \begin{cases} \left(t_{pi} < \alpha_i < \infty\right)^{y_{pi}}\left(-\infty < \alpha_i \leq t_{pi}\right)^{1-y_{pi}} & \text{if } \theta_p > 0 \\ \left(-\infty < \alpha_i < t_{pi}\right)^{y_{pi}}\left(t_{pi} \leq \alpha_i < \infty\right)^{1-y_{pi}} & \text{if } \theta_p < 0 \end{cases} \end{aligned} \tag{8}$$

where

$$t_{pi} \equiv \frac{\delta_i - z_{pi}}{\theta_p} \quad . \tag{9}$$

The indicator functions depend on the sign of θ_p because we divide by θ_p on both sides of the inequality sign in (8). The support for α_i is a product over persons. If $\theta_p > 0$,

$$l_p = \begin{cases} t_{pi} & \text{if } y_{pi} = 1 \\ -\infty & \text{if } y_{pi} = 0 \end{cases} \quad \text{and} \quad h_p = \begin{cases} \infty & \text{if } y_{pi} = 1 \\ t_{pi} & \text{if } y_{pi} = 0 \end{cases} \quad .$$

If $\theta_p < 0$, then

$$l_p = \begin{cases} -\infty & \text{if } y_{pi} = 1 \\ t_{pi} & \text{if } y_{pi} = 0 \end{cases} \quad \text{and} \quad h_p = \begin{cases} t_{pi} & \text{if } y_{pi} = 1 \\ \infty & \text{if } y_{pi} = 0 \end{cases} \quad .$$

The support of θ_p: Calculating the support for θ_p is very similar to calculating the support of α_i. The difference is that here we have a product over items. Let

$$t_{pi} \equiv \frac{\delta_i - z_{pi}}{\alpha_i} \quad . \tag{10}$$

If $\alpha_i > 0$, then

$$l_i = \begin{cases} t_{pi} & \text{if } y_{pi} = 1 \\ -\infty & \text{if } y_{pi} = 0 \end{cases} \quad \text{and} \quad h_i = \begin{cases} \infty & \text{if } y_{pi} = 1 \\ t_{pi} & \text{if } y_{pi} = 0 \end{cases} \quad .$$

If $\alpha_i < 0$,

$$l_i = \begin{cases} -\infty & \text{if } y_{pi} = 1 \\ t_{pi} & \text{if } y_{pi} = 0 \end{cases} \quad \text{and} \quad h_i = \begin{cases} t_{pi} & \text{if } y_{pi} = 1 \\ \infty & \text{if } y_{pi} = 0 \end{cases} \quad .$$

In practice, we consider each interval in (7) separately and increase (decrease) the lower bound (upper bound) of the intersection, each time we encounter an interval with a higher lower bound (lower upper bound). This is illustrated with the following *pseudo-code* description of an algorithm to determine the truncation constants for the full conditional of θ_p:

$l = -\infty$
$h = \infty$
FOR $i = 1$ to the number of items
$t_{pi} = \frac{\delta_i - z_{pi}}{\alpha_i}$
IF $y_{pi} = 0$
IF $t_{pi} < h$ and $\alpha_i > 0$ then $h = t_{pi}$
IF $t_{pi} > l$ and $\alpha_i < 0$ then $l = t_{pi}$
IF $y_{pi} = 1$
IF $t_{pi} > l$ and $\alpha_i > 0$ then $l = t_{pi}$
IF $t_{pi} < h$ and $\alpha_i < 0$ then $h = t_{pi}$
END

It is clear that the DA-T Gibbs sampler stops if any of the intersections is empty. In the next paragraph it is shown that this will never happen.

Could any of the Intersections be Empty?

For any parameter values at the jth iteration, we generate latent data such that

$$\prod_i \prod_p \left[(z_{pi}^{(j+1)} > \delta_i^{(j)} - \alpha_i^{(j)}\theta_p^{(j)})^{y_{pi}} (z_{pi}^{(j+1)} \leq \delta_i^{(j)} - \alpha_i^{(j)}\theta_p^{(j)})^{1-y_{pi}} \right] = 1 \quad .$$

This means that, at this point, we are inside the support of the posterior. Then, we draw, say, δ_i from

$$f(\delta_i | \dots) \propto \left[\prod_p (z_{pi}^{(j+1)} + \alpha_i^{(j)}\theta_p^{(j)} > \delta_i)^{y_{pi}} (z_{pi}^{(j+1)} + \alpha_i^{(j)}\theta_p^{(j)} \leq \delta_i)^{1-y_{pi}} \right] f(\delta_i) \quad .$$

Since the term within square brackets is one for $\delta_i = \delta_i^{(j)}$, it follows that the support of the full conditional is not empty. The same is true for the other parameters. It follows that none of the intersections can be empty.

Sampling from a Truncated Distribution

Let X denote a random variable with distribution function F. We wish to generate a realization of X under the condition that X takes values in the

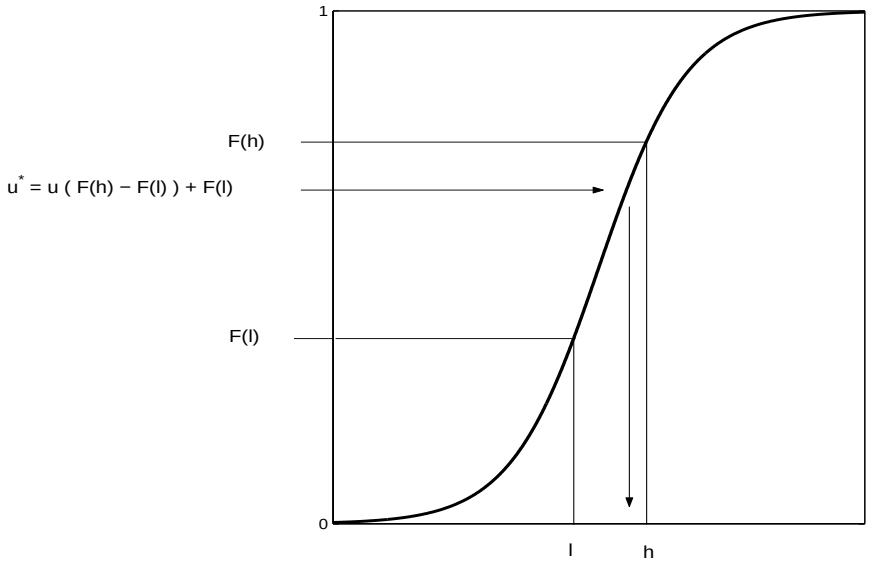

Figure 6. Simulating from a truncated distribution

interval $(l < X < h)$. Figure 6 illustrates how this may be done. First, we draw u from the uniform distribution. We then transform u to

$$u^* = \{u\,[F(h) - F(l)] + F(l)\}$$

which lies in the interval from $F(l)$ to $F(h)$. The value $F^{-1}(u^*)$ is a realization of the truncated variable.

Estimating Under Restrictions

Researchers often hold prior ideas about the parameters that take the form of order restrictions on the parameters. They may, for instance, believe item 1 to be easier than item 2. Thus, the prior density becomes

$$f(\theta, \delta, \alpha)(\delta_1 < \delta_2) \quad .$$

Each such restriction is added to the range restrictions of the full conditionals.

Handling Incomplete Designs

In applications, the design of the study is often *incomplete*. This means that only a subset of the available items is administered to each person, and no responses are observed for items that were not administered. To adapt the Gibbs sampler to handle data collected in an incomplete design we need only ignore, for each person, the items that were not administered.

Linear Logistic Test Models

In this section we demonstrate how the DA-T Gibbs sampler for the 2PL is adapted to estimate the parameters of the *Linear Logistic Test Model (LLTM)* (Fischer, 1995 and references therein).

Assume that the Rasch model is valid. That is, all discrimination parameters are unit constants. The LLTM specifies each item difficulty parameter as a linear combination of so-called basic parameters:

$$\delta_i(\eta) = q_{i1}\eta_1 + q_{i2}\eta_2 + \cdots + q_{ik}\eta_k \quad . \tag{11}$$

For ease of presentation, we assume that η_j refers to the difficulty of a mental operation, and q_{ij} to the number of times this operation is required for the i-th item. Thus, the weights q_{ij} are non-negative integers.

The DA-T Gibbs sampler for the LLTM differs very little from that of the Rasch model. Instead of sampling the item difficulties we now sample the basic parameter. If we replace, in the DA-T posterior of the Rasch model (6), δ_i by $\delta_i(\eta)$, it is easy to derive that

$$f(\eta_j|\ldots) \propto f(\eta_j) \left[\prod_p \prod_{i:q_{ij}\neq 0} (\eta_j < t_{pi})^{y_{pi}} (\eta_j \geq t_{pi})^{1-y_{pi}} \right] \quad ,$$

where $\prod_{i:q_{ij}\neq 0}$ denotes the product over all items that require the j-th mental operation, and

$$t_{pi} \equiv \frac{z_{pi} + \theta_p - \sum_{h\neq j} q_{ih}\eta_h}{q_{ij}} \quad .$$

The further specification of the DA-T Gibbs sampler is only marginally different from the DA-T Gibbs sampler for the 2PL.

For illustration, we analyse a small data set that was published by Rost (1996, pp. 99-100).[4] The data set consists of the responses of 300 persons to five geometrical analogy items. Rost (1996, section 3.4) considers the following weights appropriate for these five items:

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 1 & 1 \\ 2 & 1 \\ 2 & 2 \end{pmatrix}$$

Thus, $q_{21} = 2$, $q_{32} = 1$, etc. In this case, the weights are such that the basic parameters are uniquely determined (see Fischer, 1995).

We assume that the persons are a simple random sample from a normal population with mean μ and variance σ^2. The population parameters are estimated with the other parameters. The full conditional of μ is

$$f(\mu|\dots) \propto f(\mu) \left[\prod_p \prod_i (\mu > t_{pi})^{y_{pi}} (\mu \leq t_{pi})^{1-y_{pi}} \right] \quad ,$$

where $t_{pi} \equiv \sum_j q_{ij}\eta_j - z_{pi} - \eta_p\sigma$, and $\eta_p = (\theta_p - \mu)/\sigma$. The full conditional of σ is

$$f(\sigma|\dots) \propto f(\sigma) \left[(\sigma > 0) \prod_p \prod_i (\eta_p\sigma > \delta_i - z_{pi})^{y_{pi}} (\eta_p\sigma \leq \delta_i - z_{pi})^{1-y_{pi}} \right] \quad .$$

The other full conditionals are unchanged, but θ_p is replaced by $\eta_p\sigma + \mu$. The details are in Maris and Maris (2002, section 2.3.4).

We do not presume to know very much about the parameters and use zero-mean logistic priors with a large variance. After a *burn-in period* of $200,000$ iterations we had the program run for a few days to do several million iterations. The posterior means and standard deviations are in the following

[4] Previous (non-Bayesian) analyses on the same data are reported by Rost (1996), and Bechger, Verstralen, and Verhelst (2002, section 6).

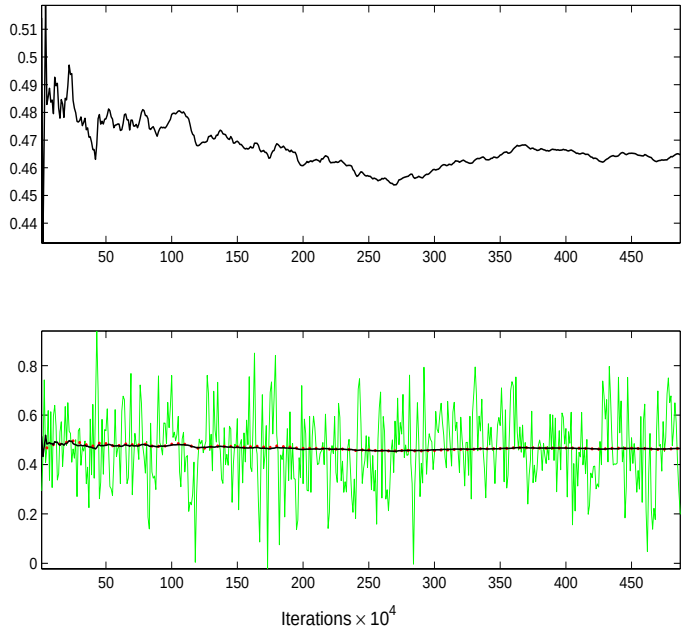

Figure 7. Summary of sampled values of the first basic parameter following a burn-in period. The first plot shows the running mean over iterations. The second plot shows sampled values. The line running through the sampled values is again the running mean.

table:

	η_1	η_2	μ	σ
posterior mean	0.463	0.969	1.361	1.993
posterior stand. dev.	0.149	0.103	0.270	0.157

The first mental operation was more difficult than the second in over 99% of the sampled values.

Figure 7 shows two plots of the running mean of the sampled values of η_1. The upper drawing suggests that the chain has not converged after the burn-in period. After an initial phase of erratic behaviour, the running mean is seen to move downwards stabilizing after about $3,500,000$ iterations. In the lower plot, however, it is seen that the variation in the running mean is negligible on the scale of the sampled values.

2PL Mixture IRT Models

A *2PL Mixture Model* (2PLMM) is an IRT model that can be written as:

$$P(Y_{pi} = j|\theta, \lambda) = \sum_s P(Y_{pi} = j|\mathbf{S} = \mathbf{s}, \lambda_{y|\mathbf{s}}) P(\mathbf{S} = \mathbf{s}|\theta, \lambda_s) \quad ,$$

where:

1. $\mathbf{S} = (S_1, \ldots, S_k)$ denotes a vector of discrete latent item responses.
2. $Y_{pi}|\mathbf{S} = \mathbf{s}$ follows a multinomial distribution.
3. $P(\mathbf{S} = \mathbf{s}|\theta, \lambda_s)$ is the likelihood of k locally independent 2PL items.
4. θ may be multi-dimensional.

2PLMMs are defined by restrictions on the distribution of Y_{pi} given $\mathbf{S} = \mathbf{s}$. Consider, for example, the 3PL. In the 3PL, $k = 1$, and

$$S = \begin{cases} 1 \text{ if a person knows the correct answer} \\ 0 \text{ if he doesn't know the correct answer} \end{cases} \quad .$$

Consequently,

$$P(Y_{ip} = 1|S = 1, \lambda_{y|s}) = 1 \quad \text{and} \quad P(Y_{ip} = 1|S = 0, \lambda_{y|s}) = \lambda_{y|s} \quad .$$

In *latent response models* (Maris, 1995), θ is multi-dimensional but the probabilities $P(Y_{pi} = j|S = s)$ are known and equal to zero or one. An example is the conjunctive Rasch model (see Maris & Maris, 2002, section 2.3.2).

The DA-T Gibbs sampler for the 2PL can be used to build a Gibbs sampler for any 2PLMM. Specifically, at each iteration we draw a sample from the posterior

$$f(\theta, \lambda, \mathbf{s}|\mathbf{y}) \propto f(\theta, \lambda, \mathbf{s}, \mathbf{y}) \quad .$$

in three *steps*:

1. Generate latent discrete item responses from $f(\mathbf{s}|\theta, \lambda, \mathbf{y})$.
2. Generate θ and λ_s from $f(\theta, \lambda_s|\mathbf{s})$.
3. Generate $\lambda_{y|s}$ from $f(\lambda_{y|s}|\mathbf{s}, \mathbf{y})$.

Due to LI, step 1 entails generating independent responses to each of k 2PL items for each of the persons. Step 2 can be done using the DA-T Gibbs sampler for the 2PL. Step 3 is the most complicated step. It is relatively simple if the prior of $\lambda_{y|s}$ is taken to be a truncated Dirichlet distribution because this implies that the corresponding full conditional is also a truncated Dirichlet distribution. In the 3PL, for instance, the full conditional of the guessing parameter would then be a truncated β-distribution. In latent response models, step 3 is unnecessary because $\lambda_{y|s}$ is known. As an illustration, we construct a Gibbs sampler for the Nedelsky model.

The Nedelsky Model for Multiple-Choice Items

Consider a multiple-choice (MC) item i with $J_i + 1$ options arbitrarily indexed $0, 1, \ldots, J_i$. For convenience, 0 indexes the only correct alternative. The *Nedelsky Model* (NM) is based upon the idea that a person responds to a MC question by first eliminating the incorrect answers (or *distractors*) he recognizes as wrong and then guesses *at random* from the remaining answers.

The probability that wrong answer j is recognized as *wrong* by a respondent with ability θ_p is modelled as a 2PL. That is, for $j = 1, \ldots, J_i$,

$$P(S_{ij} = 1|\theta_p) = \frac{\exp(\alpha_i\theta_p - \delta_{ij})}{1 + \exp(\alpha_i\theta_p - \delta_{ij})} \quad ,$$

where S_{ij} denotes a random variable that indicates whether alternative j is recognized to be wrong. Thus we may think of each distractor as a 2PL item. A "correct" answer is produced if the distractor is seen to be wrong. We will now assume that the discrimination parameter is positive.

Define a *latent subset* $\mathbf{S}_i$ by the vector $(0, S_{i1}, \ldots, S_{iJ_i})$. Assuming independence among the options given θ, the probability that a subject with ability θ_p chooses any latent subset $\mathbf{s}_i$ is given by

$$\begin{aligned} P(\mathbf{S}_i = \mathbf{s}_i|\theta_p) &= \prod_{j=1}^{J_i} \frac{\exp(\alpha_i\theta_p - \delta_{ij})^{s_{ij}}}{1 + \exp(\alpha_i\theta_p - \delta_{ij})} \\ &= \frac{\exp\left(\alpha_i\theta_p s_i^+ - \sum_{j=1}^{J_i} s_{ij}\delta_{ij}\right)}{\prod_{j=1}^{J_i}\left[1 + \exp(\alpha_i\theta_p - \delta_{ij})\right]} \quad , \end{aligned}$$

where $s_i^+ \equiv \sum_{j=1}^{J_i} s_{ij}$ denotes the number of distractors that are recognized as wrong.

Once a latent subset is chosen, a respondent guesses *at random* from the remaining answers. Thus, the conditional probability of responding with option j to item i, given latent subset $\mathbf{s}_i$, is given by:

$$P(Y_i = j|\mathbf{S}_i = \mathbf{s}_i) = \frac{1 - s_{ij}}{\sum_{h=0}^{J_i}(1 - s_{ih})} \quad ,$$

where $\sum_{h=0}^{J_i}(1 - s_{ih})$ denotes the number of alternatives to choose from.

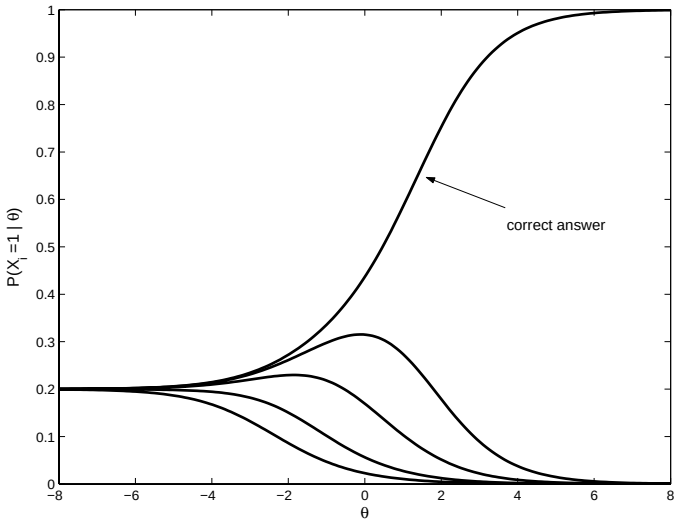

Figure 8. The item response function $P(Y_i = j|\theta)$ (with $\alpha_i > 0$) against θ for a Nedelsky item with five categories.

Combining the two stages of the response process, we find that the conditional probability of choosing option j with item i is equal to

$$P(Y_i = j|\theta_p) = \sum_{\mathbf{s}_i} \frac{1 - s_{ij}}{\sum_{h=0}^{J_i}(1 - s_{ih})} P(\mathbf{S}_i = \mathbf{s}_i|\theta_p) \quad .$$

Figure 8 shows a plot of these probabilities for an item with five categories. Note that

$$\lim_{\theta \to -\infty} P(Y_i = j|\theta) = \frac{1}{J_i + 1} \quad \text{(for } j = 0, \ldots, J_i\text{)}.$$

In fact, if an item has only two answer categories (wrong and correct), the NM equals the 3PL with the parameter $\lambda_{y|s}$ in the latter model fixed at $\frac{1}{2}$.

We will now derive a DA-T Gibbs sampler for the NM. Let $\mathbf{s}$ denote the latent subsets, and $\mathbf{s}_{ip}$ the latent subset of respondent p on answering the ith item. The vector θ contains the abilities, and the vector δ the parameters of the items. The parameters of item i are denoted by $\delta_i = (\alpha_i, \delta_{i1}, \ldots, \delta_{iJ_i})$.

We proceed by drawing a sample from $f(\theta, \delta, \mathbf{s}|\mathbf{y})$ and then ignore the latent subsets. To this aim, we consider two full conditionals: $f(\theta, \delta|\mathbf{y}, \mathbf{s}) = f(\theta, \delta|\mathbf{s})$ and $f(\mathbf{s}|\theta, \delta, \mathbf{y})$, and repeat the following steps:

1. Draw latent subsets from $f(\mathbf{s}|\theta, \delta, \mathbf{y})$.
2. Draw θ and δ from $f(\theta, \delta|\mathbf{s})$ using the Gibbs sampler for the 2PL.

Using LI and Bayes theorem it is seen that,

$$\begin{aligned} f(\mathbf{s}|\theta, \delta, \mathbf{y}) &= \prod_p \prod_i \frac{P(y_{pi}|\mathbf{s}_{ip})P(\mathbf{s}_{ip}|\theta_p, \delta_i)}{\sum_{\mathbf{s}_i} P(y_{pi}|\mathbf{s}_i)P(\mathbf{s}_i|\theta_p, \delta_i)} \\ &= \prod_p \prod_i P(\mathbf{s}_{ip}|\theta_p, \delta_i, y_{pi}) \quad . \end{aligned}$$

Hence, sampling from $f(\mathbf{s}|\theta, \delta, \mathbf{y})$ entails independently drawing $N_p N_I$ latent subsets. To this aim, we make a list of all 2^{J_i} subsets and calculate for each subset on the list the probability

$$P(\mathbf{s}_j|\theta_p, \delta_i, y_{pi}) = \frac{P(y_{pi}|\mathbf{s}_j)P(\mathbf{s}_j|\theta_p, \delta_i)}{\sum_{\mathbf{s}_i} P(y_{pi}|\mathbf{s}_i)P(\mathbf{s}_i|\theta_p, \delta_i)} \quad ,$$

where $j = 1, \ldots, 2^{J_i}$ and

$$P(y_{pi}|\mathbf{s}_j)P(\mathbf{s}_j|\theta_p, \delta_i) \propto \frac{1 - s_{j(y_{pi})}}{\sum_{h=0}^{J_i}(1 - s_{ih})} \exp\left(\alpha_i \theta s_j^+ - \sum_{k=1}^{J_i} s_{jk}\delta_{ik}\right) \quad .$$

With these probabilities we then choose a random subset from the list (see e.g., Ross, 2003, section 11.4).

Note that the NM has many parameters and hence a large number of persons is required to estimate the item parameters with reasonable precision. As an illustration we provide, in Figure 9, recovery plots of true values against

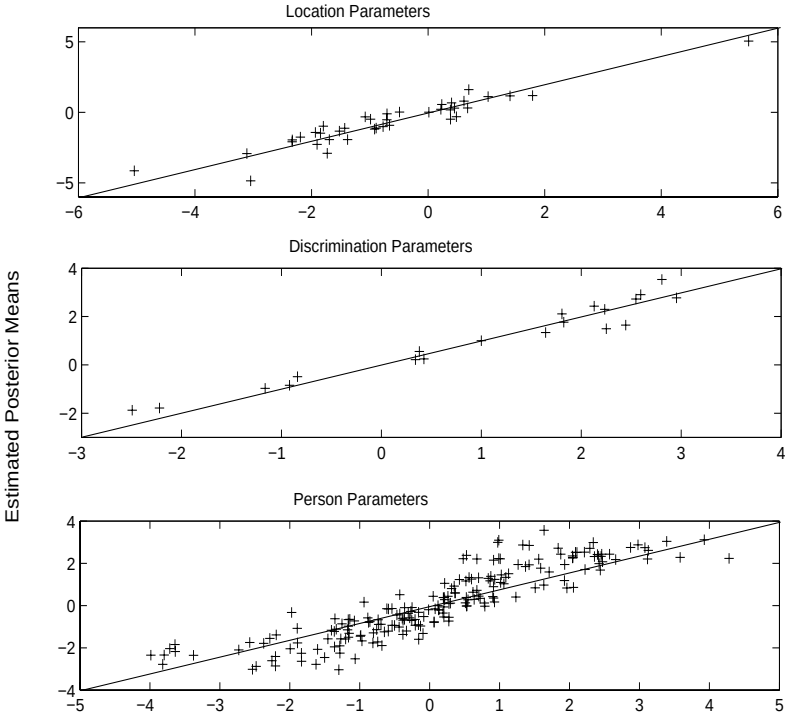

Figure 9. A typical recovery plot for an analysis with 20 items and 200 persons. Generating parameter values are on the horizontal axes. Estimated posterior means are on the vertical axes.

estimated posterior means, for a (small) data set with 20 trichotomous items and 200 persons. That is, we have simulated data under the NM, estimated the parameters, and plotted the parameter values used to generate the data against the posterior means. It is seen that recovery is not particularly good.

Discussion

In this article we have given an expository account of the DA-T Gibbs sampler for the 2PL. In addition, we have illustrated how the DA-T Gibbs sampler for the 2PL is extended to estimate models that are a special case of the 2PL (LLTM), or used as a building block to construct samplers for more

complex models (2PLMMs). Further applications can be found in Maris and Maris (2002).

The DA-T Gibbs sampler is simple to implement but may be slow to converge. Especially with large, returning applications, the algorithm may need to run longer than we can afford to wait so that it makes sense to invest time in developing and programming a more efficient (sampling) algorithm (e.g., Chib & Greenberg, 1995).

Our focus has been on Gibbs sampling. As a consequence, a number of important issues where ignored or have only been mentioned in passing. For more information on Bayesian theory and methods, we refer to general text-books, such as Bernardo and Smith (1994), Chen, Shao, and Ibrahim (2000), Gelman, Carlin, Stern, and Rubin (1995), Gill (2002), or Tanner, (1996).

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 polytomous response data. *Journal of the American Statistical Association*, **88**, 669-679.

Bechger, T. M. , Maris, G. , & Verstralen, H. H. F. M. (2005). The Nedelsky model for multiple-choice items. Chapter 10 in *New developments in categorical data analysis for the social and behavioral sciences* edited by Andries van der Ark, Marcel Croon and Klaas Sijtsma. New-York: Lauwrence Erlbaum.

Bechger, T. M. , Verstralen, H. H. F. M. , & Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika*, **67**, 123-136.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory.* New-York: Wiley.

Birnbaum, A. (1968). Some Latent Trait Models and their Use in Inferring an Examinee's Ability. pp. 395-479 In Lord, F. M. and Novick, M. R. *Statistical Theories of Mental Test Scores*. Reading:Addison-Wesley.

Butter, R. P., De Boeck, P., & Verhelst, N. D. (1998). An item response model with internal restrictions on item difficulty. *Psychometrika*, **63**, 47-63.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167-174.

Chib, S. , & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician*, **49**, 327-335.

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion) *Journal of the Royal Statistical Society*, **41**, 1-31.

Devroye, L. (1986). *Non-uniform random variate generation.* New-York: Springer-Verlag.

Fischer, G. H. (1995). The linear logistic test model. Chapter 8 in G. H. Fisher and I.W. Molenaar (Eds.). *Rasch models: Foundations, recent developments and applications*. Berlin:Springer.

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

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-472.

Gill, J. (2002). *Baysian Methods: A social and behavioral science approach.* Boca Raton: Chapman & Hall (CRS).

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, **60**, 523-547.

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

Rasch, G. (1980). *Probabilitistic models for some intelligence and attainment testst*. Chicago: The University of Chicago Press.

Ripley, B. D. (1987). *Stochastic simulation*. New-York: Wiley.

Ross, S. M. (2001). *Simulation*. 3rd Ed. New-York: Academic press.

Ross, S. M. (2003). *Introduction to probability models.* 8th Edition. New-York: Academic Press.

Rost, J. (1996). *Testtheory, testkonstruction* [Test theory, test construction]. Bern, Germany: Verlag Hans Huber.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions.* 2nd Edition. New-York: Springer.

Tanner, M. A. , & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **87**, 82-86.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, **22**, 1701-1762.