# An IRT model with a generalized Student-t link function

Caio L. N . Azevedo[a], Helio S. Migon[b]

[a] *Department of Statistics, University of Campinas,* [b] *Department of Statistical Methods, Federal University of Rio de Janeiro*

**Abstract**

In this paper we introduce a new item response theory (IRT) model with a generalized Student t-link function with unknown degrees of freedom (df), named generalized t-link (GtL) IRT model. In this model we consider only the difficulty parameter in the item response function. GtL is an alternative to the two parameter logit and probit models, since the degrees of freedom (df) play a similar role to the discrimination parameter. However, the behavior of the curves of the GtL is different from those of the two parameter models, since in GtL the curve obtained from different df's can cross each other in more than one latent trait level. The GtL model has similar proprieties to the generalized linear mixed models, such as the existence of sufficient statistics and easy parameter interpretation. Also, many techniques of parameter estimation, model fit assessment and residual analysis developed for that models can be used for the GtL model. We develop fully Bayesian estimation and model fit assessment tools through a Metropolis-Hastings step within Gibbs sampling algorithm. We consider a prior sensitivity choice concerning the degrees of freedom. The simulation study indicates that the algorithm recovers all parameters properly. In addition, some Bayesian model fit assessment tools are considered. Finally, a real data set is analyzed using our approach and other usual models. The results indicate that our model fits the data better than the two parameter models.

*Keywords:* Item response theory, generalized Student t distribution, item response function, Bayesian inference.

## 1. Introduction

Item response theory has been increasingly used to analyze psychometric data in recent years. It consists of a set of measurement models in which

the so-called latent traits and item parameters are the main ingredients (see Lord (1980) and Lord and Novick (1968)). The IRT models provide probabilities of examinees obtaining a certain score on test items that composed the test answered by these examinees. For any IRT model, the item response function (IRF), which is equivalent to the link function in generalized linear mixed models, determines the relationship between the latent traits and the item parameters. That is, IRF provides the shape of the aforementioned probabilities. Many link functions have been proposed in the literature, such as probit, logit, log-log complement, power logit and skew probit (see Bazan et al (2006) and Bazan and Bolfarine (2010)). Some works point out how sensitive the inference is when an incorrect IRF is considered (Chen et al (1999), Chen (2004) and Nagler (1994)). Among all IRFs, only the one-parameter model, based on the probit or logit link function, belongs to the class of generalized linear mixed models. In addition, this model is widely accepted and used in the psychometric literature. Also, many methods of parameter estimation and model fit assessment are available, including those developed for generalized linear mixed models. However, for many situations, this model does not fit the data properly. The two-parameter model, which is a natural extension of the one-parameter one, is more complex than this model, even though it is applicable in more situations. However, many of the mentioned interesting properties do not apply to two-parameter models. Furthermore, the two-parameter models impose that the curves of items with different discrimination parameters and same difficulty parameter cross each other only once. Also, the observed proportion of correct response approaches either 0 or 1 at a faster rate than that imposed by the probit link model. In many situations, these behaviors do not apply.

The main goal of this paper is to present a new kind of two-parameter model. A link function, based on the generalized Student-t distribution, is proposed to define a two-parameter IRF, based on the work of Kim et al (2008). Instead of the usual discrimination parameter the degrees of freedom regulates the curvature of the item characteristic curve. We show that the interpretation of the df is close to the dp, even though the curves are different. The main difference is that the ICC's (item characteristic curves) produced by our model can cross each other in more than one latent trait level. Second, the curves generated by our model approach either 0 or 1 at a faster rate than that imposed by the probit link. We developed an MCMC algorithm for estimating all parameters simultaneously. In addition, some model fit assessment tools are presented. A simulation study is performed

to assess the quality of the parameter recovery of the model and the MCMC algorithm. Furthermore, a real data set is analyzed in order to illustrate our developments. Finally, some comments about possible extensions are made.

This article is organized as follows: In Section 2 we present our model and make some comparisons with the 2PP model. The MCMC algorithm developed to fit the model is presented in Section 3. In Section 4 we perform a simulation study and in Section 5 we conduct a real data analysis. Finally, in Section 6 we present some conclusions, comments and suggestions for future research.

## 2. Model and motivation

As mentioned before, in many real data sets, the empirical curves (observed proportion of correct response at each level of the observed score) may not be suitably modeled by the two-parameter model. Figure 1 presents an example of empirical curves of different items (with approximately the same difficulty index) that cross each other more than once and that approach 0 or 1 at a faster rate than that imposed by probit link. Therefore (as we will show later), the two-parameter models may not be suitable to analyze this data set.

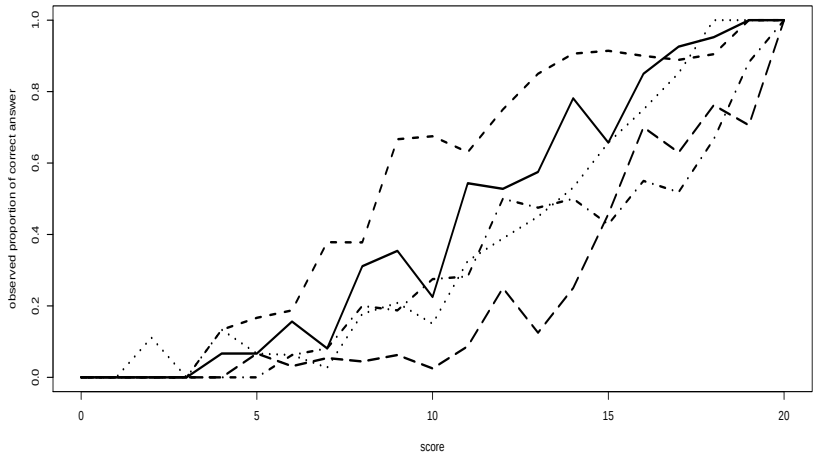


Figure 1: Empirical curves of observed proportion of correct answer by observed score (number of items correctly answered) for the real data set

To define our model we consider the situation where a set of $n$ examinees (students, patients, schools) is submitted to a measurement instrument (test, clinical evaluation, questionnaire) composed of $I$ items. Let $Y_{ij}$, the answer of examinee $j$ to item $i$, be a Bernoulli random variable, with 1 indicating a correct answer and 0 otherwise. The GtL model is given by:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(P_{ij}) \\ P_{ij} &= F_{\nu_{1i},\nu_{2i}}\left(\theta_j - b_i\right) \qquad (1) \\ \theta_j &\sim N(0,1)\,, \qquad (2) \end{aligned}$$

where $F_{(\nu_1,\nu_2)}(.)$ stands for the c.d.f of a generalized Student-t distribution function with $(\nu_1, \nu_2)$ parameters. The quantities $\nu_1$ and $\nu_2$ are the shape (degrees of freedom) and scale parameters, respectively. The density of a random variable X, $X \sim t(\nu_1, \nu_2)$ is given by

$$p(x;\nu_1,\nu_2) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\nu_1+1}{2}\right)}{\sqrt{\nu_2}\Gamma\left(\frac{\nu_1}{2}\right)} \left(1 + \frac{x^2}{\nu_2}\right)^{-\frac{1}{2}(\nu_1+1)} .$$

Notice that when $\nu_1 = \nu_2 = \nu$, we recover the standard Student t distribution, see Arellano-Valle and Bolfarine (1995). However, due to identification problems and the difficulty on interpreting $\nu_2$, we will consider $\nu_2 = 1$ Kim et al (2008). Figure 2 depicts some curves of the GtL model, for different values of $\nu_1$ but with $\nu_2 = 1$. One can see that higher the df is, the steeper is the ICC. Comparing these curves with those generated by the two-parameter probit model (see Figure 3), shows that they are different. Even though the increasing in the slope is related to the increase in either df or discrimination parameters, there are two main differences. First, in the GtL model, two curves can cross each other in more than one latent trait level. That is, the df modifies the difficulty of the items too. Second, the curves of the GtL go to zero or one slower than the curves of two-parameter model do. That is, GtL accommodates extreme probabilities more properly. As pointed out by Kim et al (2008) for a given binary response dataset, the generalized Student t-link may fit the data better than the Student-t link if the probability approaches either 0 or 1 at a faster rate than in the probit link.

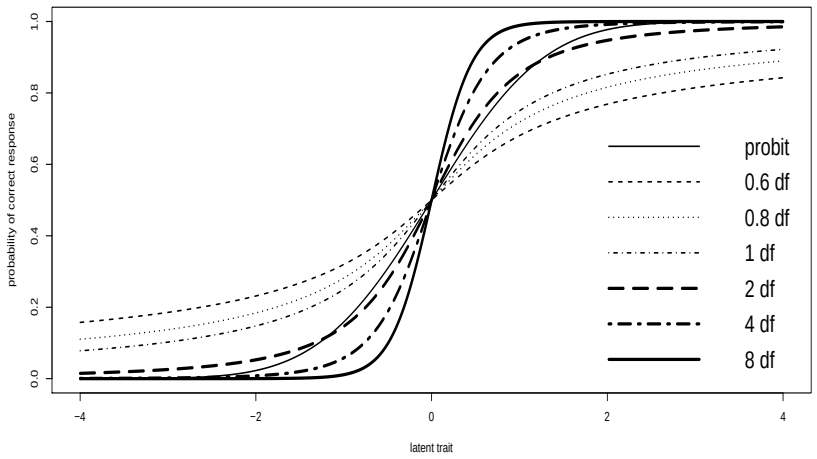


Figure 2: ICC of GtL model: latent traits versus probability of correct answer, for different values of $\nu_1$.

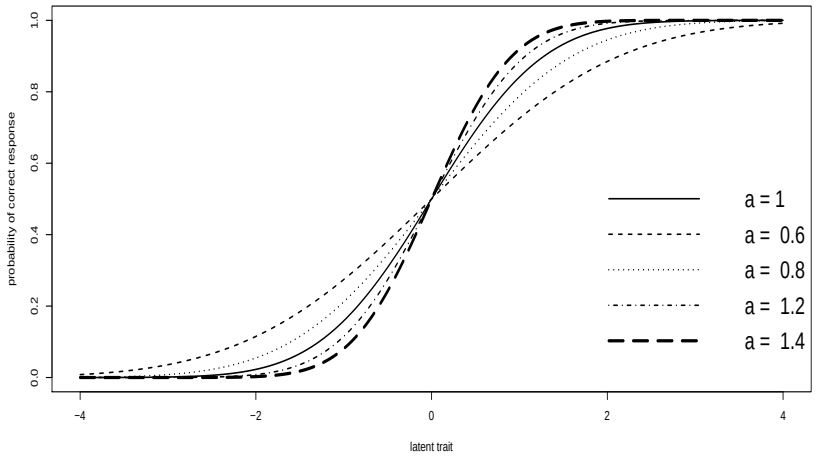


Figure 3: ICC of the two-parameter probit model: latent traits versus probability of correct answer, for different values of $\nu_1$.

## 3. MCMC estimation for the GtL model

Bayesian inference is based on the marginal posterior distributions of the parameters. Unfortunately, for the GtL and Bayesian IRT models in general, it is impossible to obtain closed-form expressions of the marginal posterior distributions. MCMC algorithms can be used to obtain samples from these posterior distributions. In order to facilitate the implementation of the MCMC algorithms, we will consider an augmented data scheme as described in Albert (1992) and a stochastic representation of the generalized Student-t distribution, given by Kim et al (2008), that is

$$\begin{aligned} Y_{ij} &= I_{(Z_{ij}\geq 0)} \\ Z_{ij}|(\theta_j, \boldsymbol{\zeta}_i, w_{ij}) &\sim N\left(\theta_j - b_i, W_{ik}^{-1}\right) \\ W_{ij}|\nu_i &\sim \text{Ga}\left(\frac{\nu_i}{2}, \frac{1}{2}\right). \end{aligned}$$

On the other hand, in many applications ignorable missing data are commonly observed. Following, for example, Azevedo et al (2011), they can be easily accommodated by considering an (observable) indicator variable:

$$V_{ij} = \begin{cases} 1, & \text{observed response of examinee } j \text{ on item } i \\ 0, & \text{missing response} \end{cases}$$

Under the usual assumptions of dichotomous IRT models, see Azevedo et al (2011) for example, the joint conditional distribution of $(\boldsymbol{Z}, \boldsymbol{W}), \boldsymbol{Z} = (Z_{11}, Z_{21}, ..., Z_{In})', \boldsymbol{W} = (W_{11}, W_{21}, ..., W_{In})$ given $(\boldsymbol{\zeta}, \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{v})$ is:

$$\begin{aligned} p(\boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{y}, \boldsymbol{v}) &= \prod_{i=1}^{I}\prod_{j=1}^{n}\left\{p(z_{ij}|w_{ij}, \theta_j, b_i, y_{ij}, v_{ij})p(w_{ij}|\nu_i)\right\} \\ &\propto \prod_{i=1}^{I}\prod_{j=1}^{n}\left\{\exp\left\{-0.5v_{ij}\left(z_{ij} - \theta_j + b_i\right)^2\right\} w_{ij}^{v_{ij}\left(\frac{\nu_i}{2}-1\right)}\right. \\ &\times \left.\exp\left(-\frac{w_{ij}v_{ij}}{2}\right)\right\}. \end{aligned} \tag{3}$$

*3.1. Prior and posterior distributions*

The joint prior distribution of the latent traits is the product of the densities given in (2). For the item parameters, we follow Azevedo et al (2011) and Liu (1996), that is

$$\begin{aligned} p(\boldsymbol{b}, \boldsymbol{\nu}) &= \prod_{i=1}^{I} \{p(b_i)p(\nu_i)\} \\ &\propto \prod_{i=1}^{I} \left\{ \exp\left\{ -\frac{1}{\psi_b} \left(b_i - \mu_b\right)^2 \right\} p(\nu_i) \right\} . \end{aligned} \tag{4}$$

The choice of the prior for the degrees of freedom plays a crucial role in the Bayesian analysis of Student-t models, see Fonseca et al (2008). In this work we perform a sensitivity study concerning this choice. More specifically, we compare the results obtained by using priors commonly considered in the literature (Geweke (1993); Fernández and Steel (1999); Fonseca et al (2008); Kim et al (2008)), that is:

$$\begin{aligned} p_1(\nu_i) &\propto \nu_i^{-2} \\ p_2(\nu_i) &\propto e^{-\lambda_\nu \nu_i} \\ p_3(\nu_i) &\propto \nu_i^{r_\lambda - 1} e^{-\lambda_\nu \nu_i} \\ p_4(\nu_i) &\propto \left(\frac{\nu_i}{\nu_i + 3}\right)^{1/2} \left[ \psi'\left(\frac{\nu_i}{2}\right) - \psi'\left(\frac{\nu_i + 1}{2}\right) - \frac{2\left(\nu_i + 3\right)}{\nu_i \left(\nu_i + 1\right)^2} \right]^{1/2} \\ p_5(\nu_i) &\propto p_4(\nu_i) \left(\frac{\nu_i + 1}{\nu_i + 3}\right)^{1/2} . \end{aligned}$$

The second tends to dominate the data concerning the likelihood, see Fonseca et al (2008). The third, which is a generalization of the second prior, presents the same problem, even though this can be attenuated by choosing the hyperparameters properly. The two last priors are independency Jeffreys and Jeffreys rule priors obtained for a Student-t regression model, respectively, see Fonseca et al (2008). The first is a limit case of the independency Jeffreys prior.

Therefore, from (2), (3) and (4), the full posterior distribution is given by

$$
\begin{aligned}
p(\boldsymbol{z},\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\zeta}|\boldsymbol{y},\boldsymbol{v}) \;\propto\; & \prod_{i=1}^{I}\prod_{j=1}^{n}\Bigg\{\exp\left\{-0.5v_{ij}\left(z_{ij}-\theta_j+b_i\right)^2\right\}w_{ij}^{v_{ij}\left(\frac{\nu_i}{2}-1\right)} \\
& \times \;\exp\left(-\frac{w_{ij}v_{ij}}{2}\right)\Bigg\}\prod_{j=1}^{n}\left\{\exp\left\{-0.5\left(\theta_j\right)^2\right\}\right\} \\
& \times \;\prod_{i=1}^{I}\left\{\exp\left\{-\frac{1}{\psi_b}\left(b_i-\mu_b\right)^2\right\}p(\nu_i)\right\} \qquad (5)
\end{aligned}
$$

*3.2. MCMC algorithms*

The distribution (5) has an intractable form, independently of the prior distribution adopted for $\nu_i$. In addition, the full conditional distribution for the degrees of freedom is not known. Therefore, a full Gibbs sampling algorithm is not feasible. However, a Metropolis-Hastings within Gibbs sampling (GS) approach is, see Patz and Junker (1999). Also, this algorithm can be slightly modified by simulating the augmented variables $\boldsymbol{W}_{i.}, (\boldsymbol{W}_{i.} = (W_{i1}, ..., W_{in}))$ and the $\nu_i$ parameters jointly, through the collapsed Gibbs technique of Liu (1994). We named this algorithm Metropolis-Hastings within collapsed Gibbs sampling (CGS). In the simulation study, we compare these two algorithms. Let $\dot{()}$ denote the set of all necessary parameters. The main steps of both algorithms are summarized below.

The Metropolis-Hastings within **Gibbs sampling** scheme

1. Start the algorithm by choosing suitable initial values.
   Repeat steps 2–6:.
2. Simulate $Z_{ij}$ from $Z_{ij} \mid (.), i = 1, ..., I,\ j = 1, ..., n.$
3. Simulate $W_{ij}$ from $W_{ij}|(.), i = 1, ..., I,\ j = 1, ..., n$
4. Simulate $\theta_j$ from $\theta_j \mid (.), j = 1, ..., n.$
5. Simulate $b_i$ from $b_i \mid (.)$, i =1,...,I.
6. Simulate $\nu_i$ from $\nu_i \mid (.)$, i =1,...,I.

The Metropolis-Hastings within **collapsed Gibbs sampling** scheme

1. Start the algorithm by choosing suitable initial values.
   Repeat steps 2–5:.
2. Simulate $Z_{ij}$ from $Z_{ij} \mid (.), i = 1, ..., I,\ j = 1, ..., n.$

3. Simulate $(\boldsymbol{W}_{i.}, \nu_i)$ jointly from $(W_{i.}, \nu_i)|(.), i = 1, ..., I$.
4. Simulate $\theta_j$ from $\theta_j \mid (.), j = 1, ..., n$.
5. Simulate $b_i$ from $b_i \mid (.)$, i =1,...,I.

Notice that the assumptions stated by equations (1) and (2) are sufficient to ensure the model identification, as in Kim et al (2008). In addition, since the priors adopted for latent traits and difficulty parameters are proper, the posterior will be proper, see Gosh et al (1993).

## 4. Simulation studies

In this section, the convergence properties and the parameter recovery of the proposed model and MCMC estimation method are discussed. The performance of the two MCMC algorithms is compared as well as the sensitivity to the choice of the prior distribution for the degrees of freedom.

### *4.1. Convergence and Autocorrelation Assessment*

An important aspect of the estimation method is assessing the convergence of the MCMC iterations. Several tests of convergence have been proposed, but there is no agreement about the most suitable one, see Gamerman & Lopes (2006). Here, we investigate the convergence of the algorithm by monitoring trace plots generated by three different sets of starting values, and by evaluating Geweke's and Gelman and Rubin's convergence diagnostics.

Item responses were simulated by considering a group of 1000 examinees answering a test of 30 items. This choice provides a sample size ratio of approximately 17 (number of latent traits divided by the number of item parameters). This value can be considered reasonable to obtain accurate estimates according to De Ayala and Sava-Bolesta (1999). The values of the parameters range from (-2.0,2.0) for the difficulty parameters and (0.4,20) for the degrees of freedom. This allows having easy, medium and difficult items as well as items with low, medium and high discrimination power. The latent traits were sampled from a standard normal distribution. The results below follow from CGS with $\mu_b = 0, \psi_b = 9$ and with the df prior $p_1(.)$ using 60,000 iterations. Similar results were found with the other priors for the degrees of freedom.

Figure 4 and 5 present the trace plots of the degrees of freedom and difficulty parameters for selected items. Sampled values were stored every 30th iteration. In each plot, three different chains are plotted, which correspond

to three different sets of initial values. From a visual inspection it can be concluded that within 100 iterations each chain of simulated values reached the same area of plausible parameter value, for the difficulty parameters. The different sets of initial values did not result in visible changes in the rate of convergence. For each starting set, each MCMC chain mixed very well, which indicates that the entire area of the parameter space was easily reached. The Geweke diagnostic, based on a burn-in period of 10,000 iterations, indicated convergence of the chains of all difficulty parameters. Furthermore, the Gelman-Rubin diagnostic ranged from .99 to 1.04, for all parameters. From both the inspection of the trace plots and the convergence diagnostics it can be concluded that the MCMC chains converged after 10,000 iterations, for these parameters. A similar pattern is observed for the degrees of freedom, excluding item 23. For this item, the mixing of the chains was mildly poor. This is due to the large value of its degrees of freedom ($\nu_{23} = 20$). This pattern is observed, in general, when the true degrees of freedom are at least 20 even though the CGS provides lesser autocorrelations than GS. This difference is illustrated by Figures 6 and 7, which present the estimated autocorrelations for the aforementioned parameters. We can notice that the autocorrelations related to CGS are significantly smaller than the autocorrelations related to GS. Thus, we can conclude that the mixing of the chains obtained from CGS is better. The problem related to degrees of freedom estimation is probably due to the fact that their estimates depend on directly on two sets of latent variables $(Z, V)$. In general, this happens in the presence of latent variables (see Leon-Gonzalez (2004)). In addition, following Sahu (2002), the effective sample size (EES) and the effective sample size per second (EESs) were calculated, for all item parameters, by using the real data set (see Section 5). As described in Sahu (2002), ESS is defined for each parameter as the number of MCMC samples drawn, B, divided by the parameter's autocorrelation time, $\gamma = 1+2\sum_{k=1}^{\infty}\rho_k$, where $\rho_k$ is the autocorrelation at lag $k$. Estimation of $\gamma$ using sample correlations is problematic because fewer MCMC samples are used in estimating $\rho_k$ as $k$ increases. There are many alternatives, see Roberts (2006) for a review. Following Sahu (2002), we use a simple upper bound $(1 + \rho^*)/(1 - \rho^*)$ where $\rho^* = \max_{k\geq 1}|\rho_k|$. In many applications $\rho^* = |\rho_1|$ and we used this for our numerical example. The results, averaged over the parameters, are presented in Table 1. The efficiency (eff), which is the ratio of ESSs of collapsed MCMC by the EESs of the MCMC, of the collapsed algorithm performs slightly better than GS.

Based on these results, we decided to consider a burn-in period of 10,000

values, storing every $50th$ values and simulating $60,000$ more values after this burn-in. Thus, we estimated the marginal posteriors using 1,000 values.

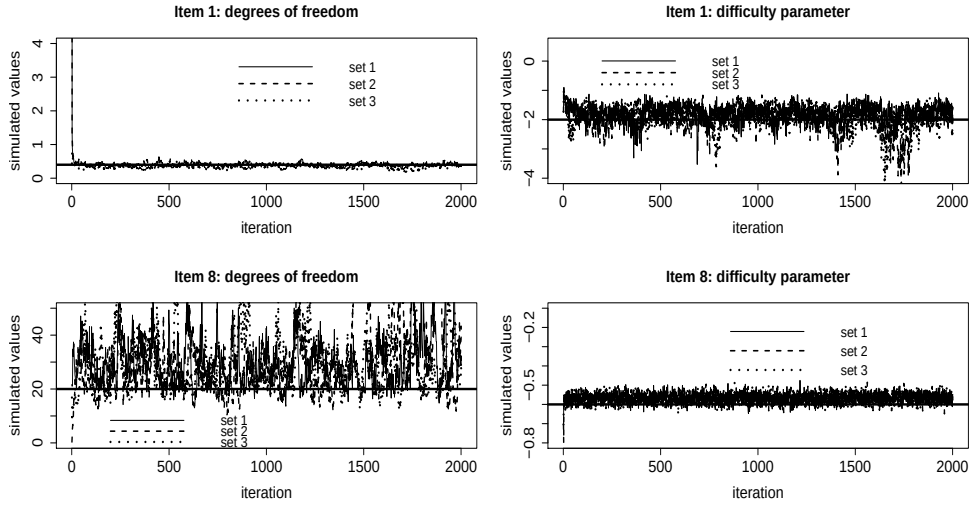


Figure 4: Trace plots of simulated values of the degrees of freedom and difficulty parameters (items 1 and 8) for different starting values

Table 1: Estimated EES of the MCMC algorithms

| **Prior** | **MCMC** | | **MCMC (collapsed)** | | **eff** |
|---|---|---|---|---|---|
| | **ESS** | **ESSs** | **ESS** | **ESSs** | |
| 1 | 1148.5 | 37.0 | 1341.9 | 43.1 | 1.2 |
| 2 | 1172.9 | 37.4 | 1359.4 | 44.6 | 1.2 |
| 3 | 1133.9 | 36.4 | 1400.7 | 47.7 | 1.3 |
| 4 | 1165.7 | 37.5 | 1440.7 | 47.2 | 1.3 |
| 5 | 1149.3 | 38.0 | 1391.4 | 45.6 | 1.2 |

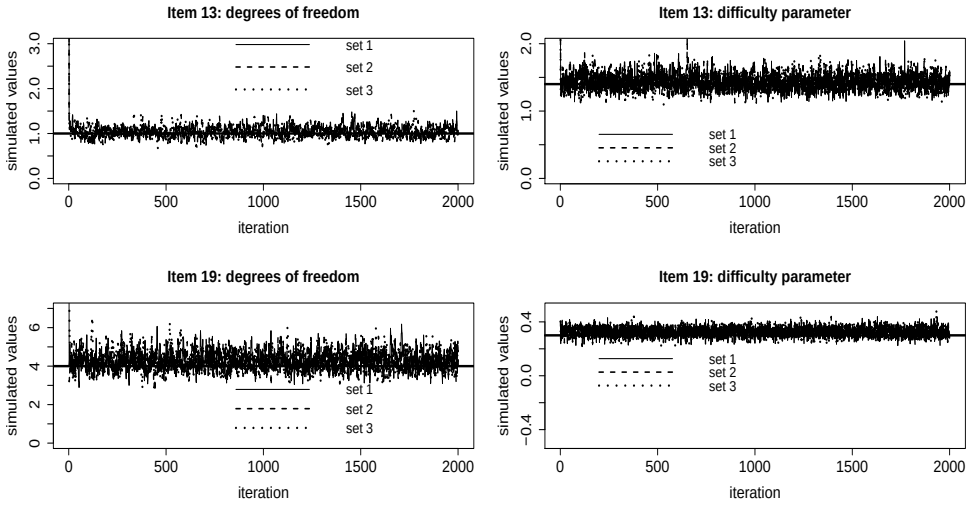


Figure 5: Trace plots of simulated values of the degrees of freedom and difficulty parameters (items 13 and 19) for different starting values

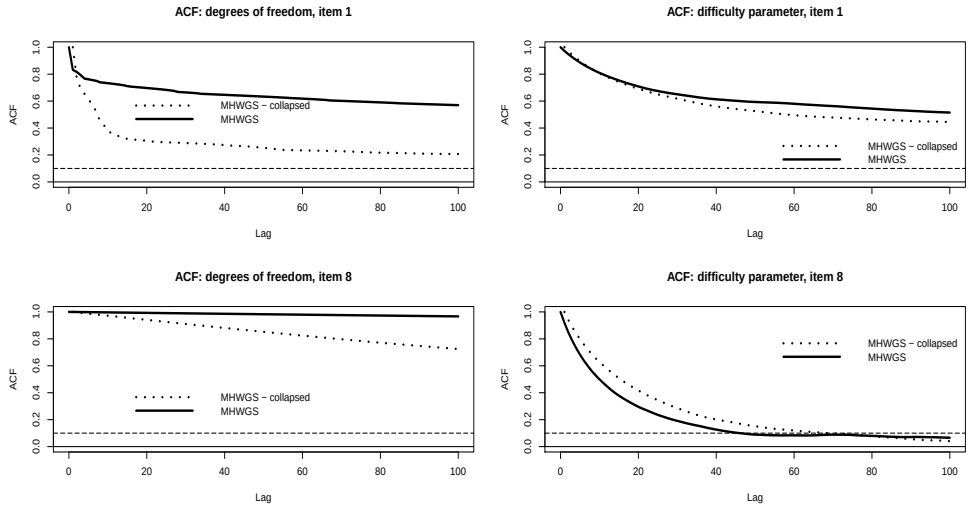


Figure 6: Estimated autocorrelations of the MCMC chains the degrees of freedom and difficulty parameters (items 1 and 8) for GS and CGS (collapsed)

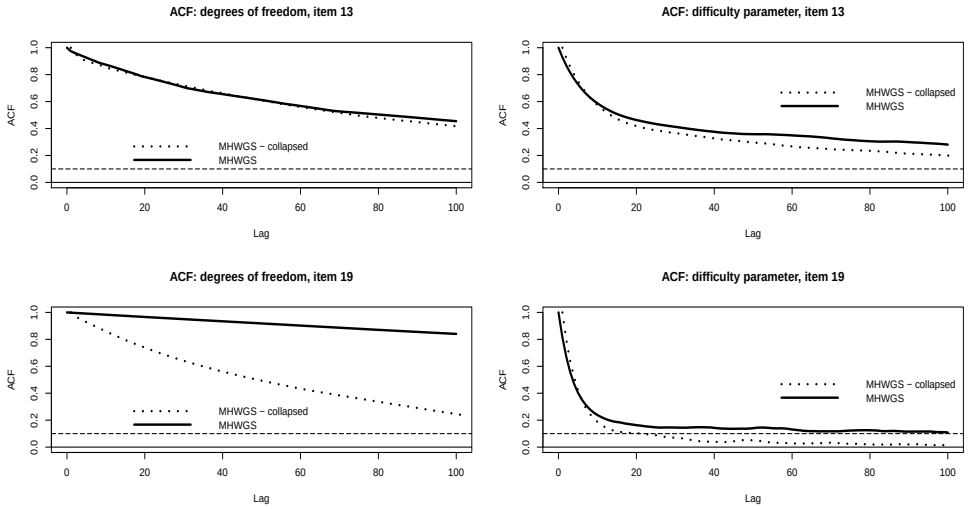


Figure 7: Estimated autocorrelations of the MCMC chains the degrees of freedom and difficulty parameters (items 13 and 19) for MHWGS and CGS (collapsed)

*4.2. Parameter Recovery Study*

The results presented in the last section showed better performance of CGS in terms of convergence. In this section we present a simulation study of prior sensitivity concerning the CGS algorithm. We do not present the results related to exponential prior $p_2(.)$, since that they are substantially worse than the others. Two factors are considered (with the levels within parenthesis), number of examinees (NE) (500,1000) and number of items (NI) (20,30). Thus, we have four situations produced by crossing the four levels. For each one of these situations, we generated a total of R = 10 replicas (that is, ten response sets). The values of the difficulty parameter and degrees of freedom were chosen in order to have from easy to difficult items as well as items with low, medium and high discrimination power. For each one of these data sets and considering each one of the prior distributions for $\nu$, the CGS was used to estimate all parameters. To compare the results of the five algorithms, we consider the square root of the mean square error (RMSE) and the absolute value of the relative bias (AVRB) based on the mean of the ten sets of parameter estimates, see Azevedo et al (2011).

By inspecting Figures 9 and 8, we can conclude that, in general, the results obtained by using the Jeffreys-rule prior are slightly better than the

others. In addition, the estimates tend to be more accurate as the ratio (NE/NI) increases. These results are consistent with those obtained by De Ayala and Sava-Bolesta (1999) and DeMars (2003). The same pattern can be observed for the AVRB (see Figures 11 and 10). In addition, the results related to latent traits (not presented) showed the same conclusions.

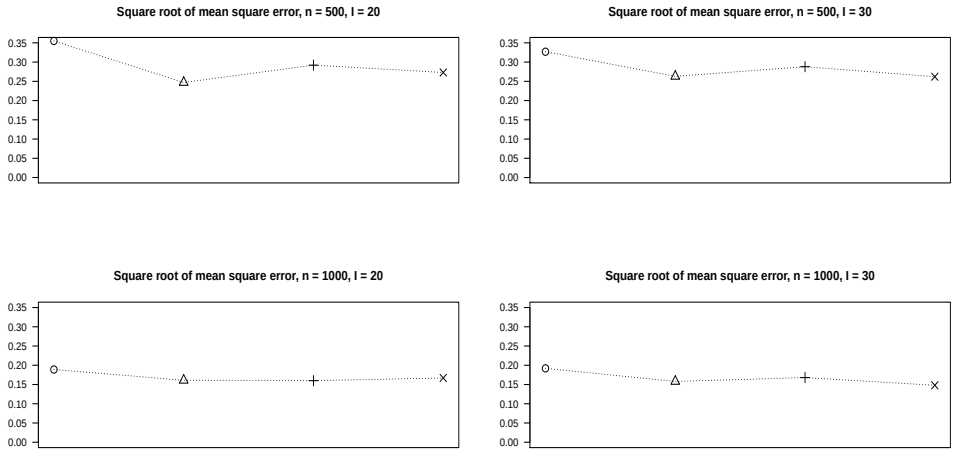


Figure 8: Root mean square error of the estimates of the difficulty parameters of ten replicated data sets for different priors: ∘ improper prior; △ gamma prior; + Jeffreys prior; × Jeffreys-rule prior.

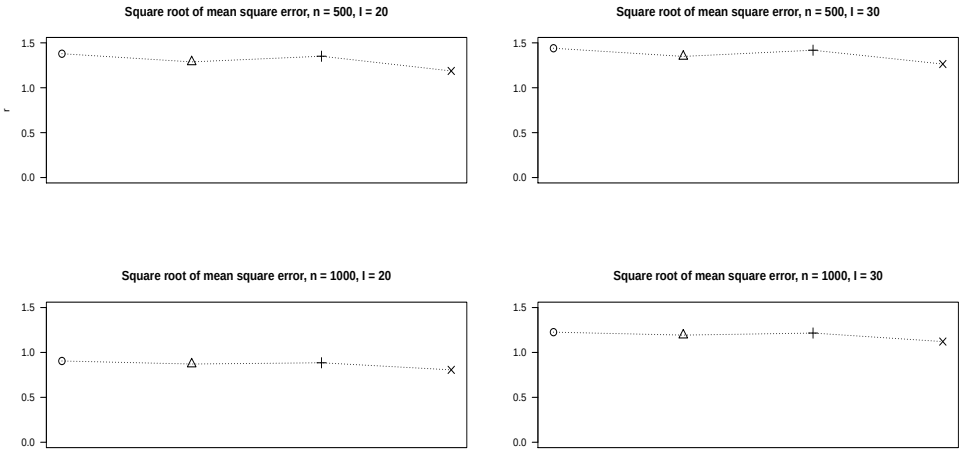


Figure 9: Root mean square error of the estimates of the degrees of freedom across ten replicated data sets for different priors: $\circ$ improper prior; $\triangle$ gamma prior; $+$ Jeffreys prior; $\times$ Jeffreys-rule prior.

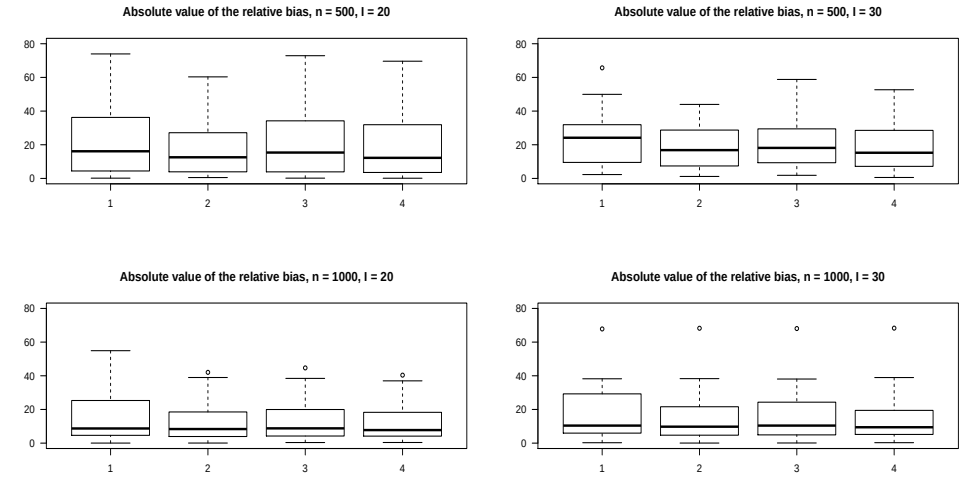


Figure 10: Absolute value of the relative bias of the estimates of the difficulty parameters across ten replicated data sets for different priors: 1 - improper prior; 2 - gamma prior; 3 - Jeffreys prior; 4 - Jeffreys-rule prior.

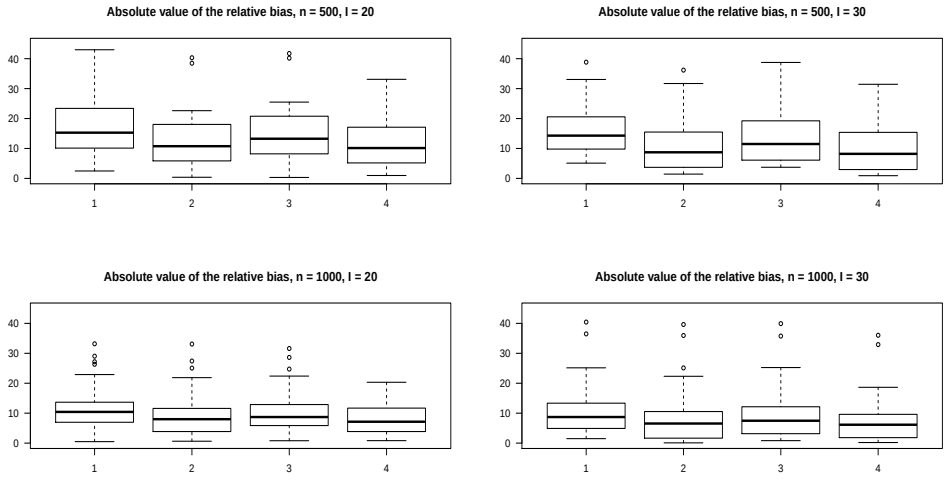


Figure 11: Absolute value of the relative bias of the estimates of the degrees of freedom across ten replicated data sets for different priors: 1 - improper prior; 2 - gamma prior; 3 - Jeffreys prior; 4 - Jeffreys-rule prior..

## 5. Real data analysis

The data set is draw from a major study related to PDE (Scholar Development Program) created by the Brazilian government. It is a program that aims to improve the teaching quality and the general structure (classrooms, libraries, informatics laboratories, etc.) in Brazilian public schools. It was implemented in 400 schools in different Brazilian states. The Brazilian Government also aimed to compare schools included in the program with schools not included, concerning mathematics and Portuguese, during five years (from fourth to eighth grade). This study was conducted from 1999 to 2003. In 1999 a sample of 158 public schools was drawn, 55 in PDE and 103 not. The sample was spread over six Brazilian states, two in each selected region (North, Northeast, and Midwest). The schools had at least 200 students enrolled in daytime, were located in urban zones and served students through eighth grade. In the baseline there were a total of 12,580 students. From 2000 to 2003 the cohort was composed of the students from the original sample who had advanced to the fifth grade and stayed at the same school. The new students enrolled in the fifth grade (coming from other schools) and

those that missed the tests in the former grade but took the current test are the second cohort, which was followed in the four subsequent years and so on. That is, the longitudinal design allowed dropouts and inclusions along the time points. In addition, several social-cultural covariables of the students were collected. In each year, one test per subject (math and Portuguese) was administered to the students.

The subset of the data that was analyzed consists of a sample of 1,500 students drawn from the fifth grade (second time point). We analyzed only the results concerning mathematics. The test was composed of 40 items.

Two models were fitted to the data. The first was the GtL with $p_5(\nu_i)$ and the second was the two-parameter probit model (2PP) as in Azevedo et al (2011). The models were fitted by using the CGS and the full Gibbs sampling algorithms, respectively. For more details concerning the use of the two-parameter model, see Azevedo et al (2011). Convergence was achieved for all item parameters, according to the statistics mentioned in Section 4.1. We did not investigate convergence for the latent traits.

Table 2 presents some model fit statistics for both models, see Spiegelhalter et al (2002) and Kass and Raftery (1995), for more details. Clearly, the GtL fitted the data better than the two-parameters model. For the two last statistics, the higher the values were the better the model fit was, with the opposite occurring for the other. Figure 12 presents the predictive and observed score distributions and qq-plot for the latent traits estimates. Some observed scores lie out the corrodibility intervals. This is probably due to the fact that the test is composed of multiple choice items. Therefore, a GtL with a guessing parameter would be more appropriate. In addition, the latent trait distribution presents heavy tails, which indicates that the use of a generalized Student-t distribution to model this distribution could appropriate. In summary, a three-parameter GtL with a generalized Student-t distribution to model the latent traits seems to be a more suitable model. However, this is beyond the scope of this article.

Figure 13 presents the posterior means and 95% HPD intervals for the difficulty parameters and degrees of freedom. Since a zero mean is assumed for the latent trait distribution, the test was difficult for these examinees. According to Figures 2 and 3, an item with a value higher than one for the degrees of freedom has a reasonable discrimination power. Thus, 26 items can be considering as having good discrimination powers. Under the two-parameters model, only 24 items can be classified as having good discrimination power ($a > 0.6$). Thus, our model, in this case, was able to extract

more information from the data than the two-parameters model, in terms of latent trait estimation.

Table 2: Statistics of model fit

| Model | $\overline{D}(\vartheta)$ | $D(\overline{\vartheta})$ | $\rho_D$ | $I\!E(AIC)$ | $I\!E(BIC)$ | predictive lik. | LPLM |
|---|---|---|---|---|---|---|---|
| GtL | 63213.0 | 61756.0 | 1456.3 | 63293.0 | 63653.0 | -31543.0 | -32616.0 |
| 2PP | 63326.0 | 61871.0 | 1454.3 | 63406.0 | 63766.0 | -31600.0 | -32682.0 |

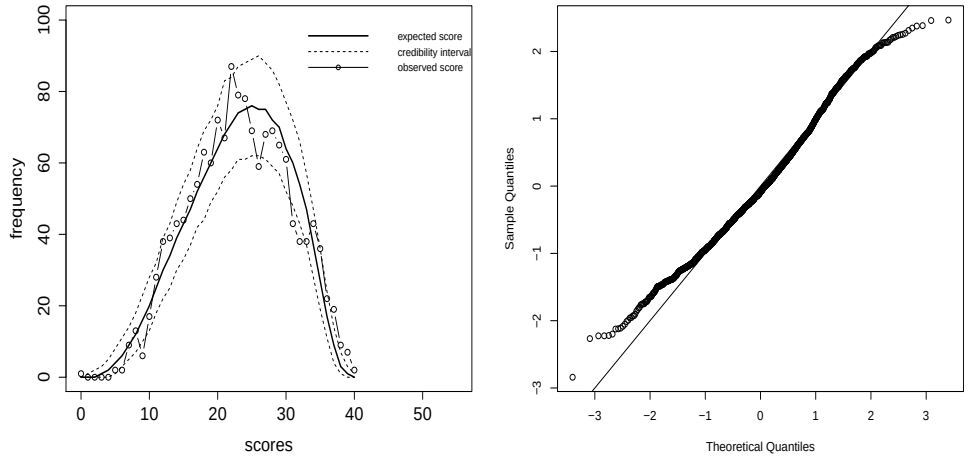


Figure 12: Predictive and observed score distributions and qq-plot for the latent traits estimates

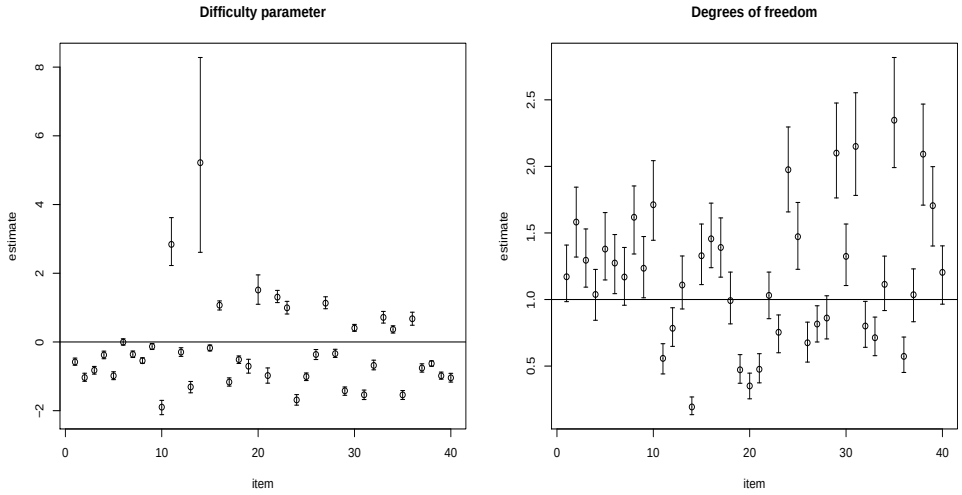


Figure 13: Difficulty parameter and degrees of freedom estimates and 95% HPD intervals

## 6. Final comments

We presented an IRT model with a link function based on the generalized Student-t distribution. This approach is an alternative to the two-parameters model, since the degrees of freedom play a similar role to the discrimination parameters. We developed an MCMC algorithm for the model fit. Such approach properly recovers all parameters, according to the simulation study. The GtL model better fit the real data set studied than the 2PP model. Also, more items were classified as having good discrimination power through our model than using the two-parameters model. In conclusion: our approach is a promising alternative to the usual ones in analyzing IRT data sets. For future research we intend to work on extensions for multiple groups and longitudinal frameworks, considering the generalized Student-t distributions for both IRF and latent trait distribution. Also, the use of other estimation methods and model fit assessment tools should be investigated. For example, the tools already developed for the GLMM could be used or adapted for the GtL model.

## 7. Acknowledgments

The authors would like to thank CENAPAD (National Center for High Performance), from Brazil, for the computational support that made the simulation studies and data analysis possible. This work was partially supported by grants from CNPq and Capes.

## References

Albert, J., (1992) Bayesian estimation of normal ogive item response curves using Gibbs sampling, *Journal of educational statistics, 17, 3, 251-269.*

Arellano-Valle, R. and Bolfarine, H. (1995) On some characterizations of the t-distribution, *Statistics & Probability Letters, 25, 1, 79-95.*

Azevedo, C. L. N., Andrade, D. F. and Fox, J.-P. (2011), Gibbs Sampling Based Estimation Procedure and Model-Fit Assessment for the Multiple-Group IRT Model, *under review, Computational Statistics & Data Analysis.*

Bazán, J. L., Branco, M. D. and Bolfarine, H. (2006), *A Skew Item Response Model*, Bayesian analysis, 861-892.

Bazán, J. L. and Bolfarine, H. (2010), *Bayesian Estimation of the Logistic Positive Exponent IRT Model*, Journal of educational and behavioral statistics, 35, 6, 693-713 .

Chen, M..-H, Dey, D. K. and Shao, Q.-M., A New Skewed Link Model for Dichotomous Quantal Response Data, *Journal of the American Statistical Association, 94, 448, 1172-1186.*

Chen, M.-H. (2004) The skewed link models for categorical response data, *In: Skew-elliptical distributions and their applications: A journey beyond normality*, Genton. M. G., ed., Boca Raton, FL: Chapman & Hall/CRC, London

Fernández, C. and Steel, M. F. J., (1999) Multivariate Student-t regression models: pitfalls and inference, *Biometrika, 86, 359-371.*

De Ayala, R. J., and Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model, *Applied Psychological Measurement, 23, 3-19.*

DeMars, C. E. (2003), Sample Size and the Recovery of Nominal Response Model Item Parameters, *Applied psychological measurement*, 27, 4, 275–288.

Fox, J.-P. (2004), Multilevel IRT assessment, *in New Developments in Categorical Data Analysis for the Social and Behavioral Sciences, van der Ark, Croon and Sijtsma eds*, Lawrence Erlbaum Associates, Inc, London.

Fonseca, T. C. O., Ferreira, M. A. R. and Migon, H. S. (2008) Objective Bayesian analysis for the Student-t regression model , *Biometrika, 95, 2, 325-333.*

Gamerman, D., & Lopes, H. (2006) *Markov chain concepts related to sampling algorithms. In: Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (Eds.), Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC, London, pp 45-57.

Geweke, J., (1993) Bayesian treatment of the independent Student-t linear model, *Journal of applied econometrics, 8, 19-40.*

Ghosh, G., Chosh, A., Chen, M.-H. and Agresti, A. (2000) Noninformative priors for one-parameter item response models, *Journal of Statistical Planning and Inference, 88, 1, 99-115.*

Kass, R. E. and Raftery, A. E. (2002), *Bayes factors*, Journal of the American Statistical Association, 90, 330, 773–795.

Kim, R. S., Chen, M.-H. and Dey, D. K., (2008) Flexible generalized t-link models for binary response data, *Biometrika, 95, 1, 93-106.*

Azevedo, C. L. N., Bolfarine H. and Andrade, D. F.,Bayesian inference for a skew-normal IRT model under the centred parameterization,*Computational Statistics & Data Analysis, 55, 353-365.*

Leon-Gonzalz, R. (2004), Data Augmentation in the Bayesian Multivariate Probit Model, *Sheffield economic research paper series, SERP number: 2004001.*

Liu, C., (1992) Bayesian robust multivariate linear regression with incomplete data, *Journal of the American statistical association, 91, 1219-1277.*

Liu, J. S., (1994) The collapsed Gibbs sampler algorithm in Bayesian computations with applications to a gene regulation problem, *Journal of the American statistical association, 89, 958-966.*

Lord, F. M., (1980) *Applications of Item Response Theory to Practical Testing Problems, first edition*, Hillsdale: Lawrence Erlbaum Associates, NJ.

Lord, F. M. and Novick, M. R., (1968) *Statistical Theories of Mental Test Scores, 592, first*, Reading: Addison-Wesley, MA.

Nagler, J. (1994), Scobit: an alternative etimator to logit and probit, *American Journal Political Science, 38, 230-255.*

A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models, *Journal of Educational and Behavorial Statistics, 24, 2, 146-178.*

Gamerman, D., & Lopes, H. (2006) *Markov Chain Monte Carlo : Stochastic simulation for bayesian inference, second edition*, Chapman & Hall/CRC, London.

Sahu, S. K., (2002) Bayesian Estimation and Model Choice in Item Response Models, *Journal of Statistical Computation and Simulation, 72, 3, 2002, 217-232.*

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linden, A. (2002), *Bayesian measures of model complexity and fit*, Journal Royal Statistical Society, 64, 3 583 – 639.