Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: www.tandfonline.com/journals/cjas20

A multiple group item response theory model with centered skew-normal latent trait distributions under a Bayesian framework

Jose R.S. Santos, Caio L.N. Azevedo & Heleno Bolfarine

To cite this article: Jose R.S. Santos, Caio L.N. Azevedo & Heleno Bolfarine (2013) A multiple group item response theory model with centered skew-normal latent trait distributions under a Bayesian framework, Journal of Applied Statistics, 40:10, 2129-2149, DOI: 10.1080/02664763.2013.807331

To link to this article: https://doi.org/10.1080/02664763.2013.807331

Journal of Applied Statistics, 2013
Vol. 40, No. 10, 2129–2149, http://dx.doi.org/10.1080/02664763.2013.807331

Taylor & Francis
Taylor & Francis Group

A multiple group item response theory model with centered skew-normal latent trait distributions under a Bayesian framework

Jose R.S. Santos[a], Caio L.N. Azevedo[a]* and Heleno Bolfarine[b]

[a]*Department of Statistics, University of Campinas, Campinas, Brazil;* [b]*Department of Statistics, University of São Paulo, São Paulo, Brazil*

(*Received 5 July 2012; final version received 17 May 2013*)

Very often, in psychometric research, as in educational assessment, it is necessary to analyze item response from clustered respondents. The multiple group item response theory (IRT) model proposed by Bock and Zimowski [12] provides a useful framework for analyzing such type of data. In this model, the selected groups of respondents are of specific interest such that group-specific population distributions need to be defined. The usual assumption for parameter estimation in this model, which is that the latent traits are random variables following different symmetric normal distributions, has been questioned in many works found in the IRT literature. Furthermore, when this assumption does not hold, misleading inference can result. In this paper, we consider that the latent traits for each group follow different skew-normal distributions, under the centered parameterization. We named it skew multiple group IRT model. This modeling extends the works of Azevedo *et al*. [4], Bazán *et al*. [11] and Bock and Zimowski [12] (concerning the latent trait distribution). Our approach ensures that the model is identifiable. We propose and compare, concerning convergence issues, two Monte Carlo Markov Chain (MCMC) algorithms for parameter estimation. A simulation study was performed in order to evaluate parameter recovery for the proposed model and the selected algorithm concerning convergence issues. Results reveal that the proposed algorithm recovers properly all model parameters. Furthermore, we analyzed a real data set which presents asymmetry concerning the latent traits distribution. The results obtained by using our approach confirmed the presence of negative asymmetry for some latent trait distributions.

Keywords: item response theory; centered skew-normal; Bayesian estimation; model identifiability; multiple group model; MCMC algorithms

1. Introduction

In psychometric research, specially in educational assessments, it is common to observe respondents (subjects) from different groups. Typically, these groups can be characterized by grade, geographic regions, gender, schools, socio-economic level and so on. Differences among the groups behavior can reflect differences among the respondents from different groups and some

*Corresponding author. Email: cnaber@ime.unicamp.br

similarity among the respondents from the same group. Therefore, it is important to take the specific group behaviors into account. Attention will be focused on applications where the number of groups is limited with a large number of respondents.

Bock and Zimowski [12] presented an item response theory (IRT) model where each group has a specific latent trait distribution. This multiple group model (MGM) has an additional set of parameters namely *multiple population parameters*, which characterize the latent population distributions. A typical assumption for parameter estimation in this model is to assume that the latent traits are random variables which follow possibly different symmetric normal distributions (with possible different population parameters). However, this assumption can be unrealistic. In many works, considering the context of one group of respondents, as in [3,11,25,33], IRT data sets, where the assumption of normality latent trait distribution is not reasonable, are presented. More recently, in the presence of multiple groups, Azevedo *et al*. [3,7] observed lack of normality for at least one of the groups. Lack of normality of the latent traits distribution is related to the presence of at least one of the following characteristics: asymmetry, heavy tails and a kurtosis different from that of the normal distribution. In the presence of multiple groups, one is more likely to observe at least one of those characteristics in the latent trait distributions than in the one-group framework. On the other hand, in the literature, there are many evidences that the lack of normality can lead to biased estimates and, consequently, to misleading inferences. From dichotomous to polytomous models, in the context of one group of respondents, works such as [2,14,15,19,20,24,36,38,39] revealed that latent trait distribution departing from normality can lead to biased estimates and such assumption has significant impact on the accuracy of estimates. Similar results, for the multiple-group framework, can be found in [3,7].

Our main goal is to extend the model proposed in [12] by modeling the latent trait distributions through group-specific skew-normal distributions under the centered parameterization. Our approach also extends the work of Azevedo *et al.* [4], concerning the number of groups, and the work of Azevedo *et al.* [3] and Bazán *et al.* [11], concerning the latent trait distributions. For Bayesian parameter estimation, we propose two Monte Carlo Markov Chain (MCMC) algorithms. The model fit assessment is made by using graphical techniques and usual statistics of model fitting. This article is organized as follows. In Section 1, we presented a literature review and the goals of this work. In Section 2, we present the model and study aspects concerning its identifiability. In Section 3, two MCMC algorithms are developed to fit the model. 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. The main conclusion is that the proposed model can be useful in dealing with multiple group IRT data where there is evidence of asymmetry of the latent trait distributions.

2. The MGM with centered skew-normal latent trait distributions

In this paper, we deal with situations where one or more different tests are administered to the respondents of each group. The tests have common items and the structure can be recognized as an incomplete block design [26], even though some (or all) tests can be identical. We will assume that each group has a reasonable number of subjects. In summary, we are dealing with a set of n respondents belonging to K groups, with n_k respondents in group k, and $n = \sum_{k=1}^{K} n_k$. The respondents within group k answer I_k items, and $\sum_{k=1}^{K} I_k < I$, where I is the total number of items.

The following notation will be introduced: θ_{jk} is the latent trait of respondent j ($j = 1, \ldots, n_k,$) belonging to group k ($k = 1, \ldots, K$), $\boldsymbol{\theta}_{.k} = (\theta_{j1}, \ldots, \theta_{jK})^\top$ is the vector of latent traits of the respondents for group k and $\boldsymbol{\theta}_{..} = (\boldsymbol{\theta}_{.1}, \ldots, \boldsymbol{\theta}_{.K})^\top$ is the vector with all latent traits; Y_{ijk} corresponds to the response of respondent j, of group k to item i ($i = 1, \ldots, I$), $\boldsymbol{Y}_{.jk} = (Y_{1jk}, \ldots, Y_{Ijk})^\top$ is the

response vector for respondent j in group k, $\boldsymbol{Y}_{..k} = (\boldsymbol{Y}_{.1k}^\top, \ldots, \boldsymbol{Y}_{.n_k k}^\top)^\top$ is the response vector of all respondents of group k, $\boldsymbol{Y}_{...} = (\boldsymbol{Y}_{..1}^\top, \ldots, \boldsymbol{Y}_{..K}^\top)^\top$ is the whole response set, and $\boldsymbol{y}_{.jk}, \boldsymbol{y}_{..k}$ and $\boldsymbol{y}_{...}$ are the corresponding observed values; $\boldsymbol{\zeta}_i$ is the vector of parameters of item i, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1^\top, \ldots, \boldsymbol{\zeta}_I^\top)^\top$ is the whole set of item parameters, $\boldsymbol{\eta}_{\theta_k}$ is the vector with the population parameters k and $\boldsymbol{\eta}_{\boldsymbol{\theta}} = (\boldsymbol{\eta}_{\theta_1}^\top, \ldots, \boldsymbol{\eta}_{\theta_K}^\top)^\top$ is the whole set of population parameters.

When modeling the grouping structure of subjects using group-specific skew-normal distributions with the centered parameterization (CP) for the latent traits, the SMGIRT (skew multiple group IRT) model can be seen as a natural extension of the two-parameter (probit) item response MGM [10,12]. That is, the full structure of the proposed model is represented by

$$\begin{aligned} Y_{ijk} \mid (\theta_{jk}, \boldsymbol{\zeta}_i) &\sim \text{Bernoulli}(P_{ijk}), \\ P_{ijk} = P(Y_{ijk} = 1 \mid \theta_{jk}, \boldsymbol{\zeta}_i) &= \Phi(a_i\theta_{jk} - b_i), \\ \theta_{jk} \mid \boldsymbol{\eta}_{\theta_k} &\sim \text{SN}_\text{C}(\mu_{\theta_k}, \sigma_{\theta_k}^2, \gamma_{\theta_k}), \end{aligned} \tag{1}$$

where $\Phi(\cdot)$ stands for the cumulative normal function. In this parameterization, the difficulty parameter $b_i = a_i b_i^*$ is a transformation of the commonly used difficulty parameter denoted as b_i^*. For more details, see [10]. In addition, $\boldsymbol{\eta}_{\theta_k} = (\mu_{\theta_k}, \sigma_{\theta_k}^2, \gamma_{\theta_k})$ and $\text{SN}_\text{C}(\mu_{\theta_k}, \sigma_{\theta_k}^2, \gamma_{\theta_k})$ stands for a skew-normal distribution under the CP (SN_C distribution), with mean μ_{θ_k}, variance $\sigma_{\theta_k}^2$ and asymmetry coefficient γ_{θ_k}. Therefore, we can see that the parameters of the SN_C distribution have straightforward interpretations, unlike the parameters of the usual skew-normal distribution [29]. In addition, the Fisher information matrix obtained through the SN_C distribution is nonsingular $\forall \gamma_{\theta_k}$ and the likelihood is well behaved, unlike the usual skew-normal distribution. For more details, see [4,5,9,29,34].

On the other hand, it is possible to prove [6,34] that the density of the skew-normal distribution under the CP is given by

$$\begin{aligned} f(\theta_{jk} \mid \boldsymbol{\eta}_{\theta_k}) &= 2\frac{\sqrt{\sigma^2}}{\sigma_{\theta_k}}\phi\left[\frac{\sqrt{\sigma^2}}{\sigma_{\theta_k}}\left(y - \mu_{\theta_k} + \frac{\sigma_{\theta_k}}{\sqrt{\sigma^2}}\mu\right)\right]\Phi\left(\lambda\left[\frac{\sqrt{\sigma^2}}{\sigma_{\theta_k}}\left(y - \mu_{\theta_k} + \frac{\sigma_{\theta_k}}{\sqrt{\sigma^2}}\mu\right)\right]\right) \\ &= 2\omega_{\theta_k}^{-1}\phi(\omega_{\theta_k}^{-1}(y - \xi_{\theta_k}))\Phi[\lambda_{\theta_k}(\omega_{\theta_k}^{-1}(y - \xi_{\theta_k}))], \end{aligned} \tag{2}$$

which corresponds to a skew-normal distribution under Henze's [21] stochastic representation with parameters ξ_{θ_k}, ω_{θ_k} and λ_{θ_k} where

$$\xi_{\theta_k} = \mu_{\theta_k} - \frac{\sigma_{\theta_k}\mu}{\sigma}, \quad \omega_{\theta_k} = \frac{\sigma_{\theta_k}}{\sigma}, \quad \mu = r\delta_{\theta_k}, \quad \sigma^2 = 1 - \mu^2, \quad \sigma = \sqrt{\sigma^2} \text{ and } \delta_{\theta_k} = \frac{\lambda_{\theta_k}}{1 + \lambda_{\theta_k}^2}. \tag{3}$$

On the other hand, in terms of centered parameters $\mu_{\theta_k}, \sigma_{\theta_k} = \sqrt{\sigma_{\theta_k}^2}$ and γ_{θ_k}, we have the following relations:

$$\begin{aligned} \xi_{\theta_k} &= \mu_{\theta_k} - \sigma_{\theta_k}\gamma_{\theta_k}^{1/3}s, \\ \omega_{\theta_k} &= \sigma_{\theta_k}\sqrt{1 + \gamma_{\theta_k}^{2/3}s^2}, \\ \lambda_{\theta_k} &= \frac{\gamma_{\theta_k}^{1/3}s}{\sqrt{r^2 + s^2\gamma_{\theta_k}^{2/3}(r^2 - 1)}}, \end{aligned}$$

$$\text{where} \quad s = \left(\frac{2}{4 - \pi}\right)^{1/3} \quad \text{and} \quad r = \sqrt{\frac{2}{\pi}}.$$

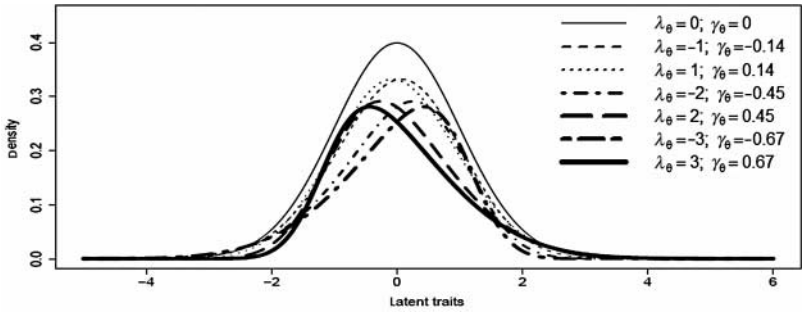

Figure 1. Skew-normal plots for different values of λ_θ (γ_θ).

It is also relevant to note that the density given by Equation (2) does not depend on the particular stochastic representation considered, either Sahu's or Henze's (see [21,32], respectively).

Figure 1 shows examples of densities given by Equation (2) for different values of λ_{θ_k} (γ_{θ_k}). The centered skew-normal ($\mathrm{SN_C}$) distribution is parameterized by the skewness coefficient (γ_θ) instead of the asymmetry parameter (λ_θ). The parameter γ_θ has a straightforward interpretation, since the closer it is to -1 or 1, the higher is the negative or the positive asymmetry of the latent traits distribution. In addition, if $\gamma_\theta \in (-0.13, 0.13)$ the distribution is considered to be symmetric. Hence, it is clear that, since the parameter space corresponding to γ_θ is a bounded set, it makes simpler its interpretation, the choice of prior distributions and kernel densities (to be used with Metropolis–Hastings algorithms).

2.1 *Model identification*

As in the MGM with symmetric normal distribution, to ensure model identification, it suffices to assume that the expectation and the standard deviation of the reference group (in this case, group 1) are fixed, for example, at zero and one, respectively. In other words,

$$\theta_{j1} \sim \mathrm{SN_C}(0, 1, \gamma_{\theta_1}),$$
$$\theta_{jk} \sim \mathrm{SN_C}(\mu_{\theta_k}, \sigma^2_{\theta_k}, \gamma_{\theta_k}), \quad k = 2, \ldots, K,$$

with a suitable linking design in terms of the administered tests. That is, the tests need to have some structure of common items. In other words, by fixing the mean and the standard deviation of the reference group and by considering some structure of common items, among the tests, the model is identified. For more details, see [3,4,34].

Therefore, the metric (scale) is defined and model (1) is identified due to the fact that model (1) is no longer invariant to location-scale transformations, since that the expected value and the standard deviation of the latent distribution of the reference group (in this case, group 1) are fixed and also due to the linking design. This ensures that the metric for the latent traits is well defined and the results related to all tests (item parameters) and groups (latent traits and population parameters) lie on the same scale. In addition, the likelihood of our model is much improved compared with the one based on the ordinary skew-normal distribution. Further details can be found in [4].

As a final comment, we can mention that in the literature, other item response functions (IRFs) such as the skew-probit, logit or log–log can be considered. To include this flexibility in response functions, a generalized MGM is defined as a mixture (indexed $l = 1, \ldots, L$) of different response functions, that is, a composite link function [22], based on different cumulative

distribution functions (indexed $h = 1, \ldots, H$). Then, the success probability is stated as

$$P_{ijk} = P(Y_{ijk} = 1 \mid \theta_{jk}, \boldsymbol{\zeta}_i, \boldsymbol{\nu}) = \sum_{l=1}^{L} \prod_{h=1}^{H} F_{lh}(\theta_{jk}, \boldsymbol{\zeta}_i, \boldsymbol{\nu}), \tag{4}$$

where $F(\cdot)$ represents a cumulative distribution function with parameters $\boldsymbol{\nu}$.

This modeling framework comprehends the well-known one-, two- and three-parameter item response models using probit, logit, log–log, Student's-t, skew probit or skew-probit link function, among others, see [2,8,11]. Note also that extensions to nominal and ordinal response data can be made by defining a different response model at level 1 of the MGM, given by Equation (4). In the same way, the MGM for mixed response data will contain response models for discrete binary and polytomous response data.

Next we present a Bayesian approach for parameter estimation.

3. Bayesian estimation and MCMC algorithms

Bayesian inference is based on the posterior distribution of the model parameters, which is proportional to the product of the likelihood and a prior distribution. In IRT, very often, it is not possible to obtain the posterior distributions, analytically. However, MCMC algorithms have been successfully used in providing empirical approximations for them, under some conditions [17]. Concerning the augmented data and indicator matrices, presented ahead, we will follow [3] in such implementation. An augmented data scheme is introduced to sample continuous normally distributed item response data, denoted as $\mathbf{Z}$, given discrete observed item response data, denoted as $\mathbf{y}$. According to [1], we have

$$Z_{ijk} \mid (\theta_{jk}, \boldsymbol{\zeta}_i, Y_{ijk}) \sim N(a_i\theta_{jk} - b_i, 1), \tag{5}$$

where Y_{ijk} is the indicator of Z_{ijk} being greater than zero.

To handle an incomplete block design, an indicator variable $\boldsymbol{I}$ is introduce which defines the set of administered items according to the design. For each administered item response, the corresponding information is recorded. This indicator variable is described by

$$I_{ijk} = \begin{cases} 1, & \text{if item } i \text{ is administered for respondent } j \text{ of group } k, \\ 0, & \text{if missing by design.} \end{cases}$$

The indicator matrix $\boldsymbol{I}$ describes the patterns of missing responses that are deliberately allowed to be missing. These missing data are missing by design.

In addition, an indicator variable can be defined to describe the missingness due to uncontrolled events as nonresponse or errors in recoding data. The missing data indicator variable is defined as

$$V_{ijk} = \begin{cases} 1, & \text{if response of respondent } j \text{ of group } k \text{ on item } i \text{ is recorded,} \\ 0, & \text{if response is missing.} \end{cases}$$

Indicator variables $\boldsymbol{V}_{\ldots} = (V_{111}, \ldots, V_{In_KK})$ refer to observed data that could be missing.

In case of missing at random (MAR), the missing indicator matrix, $\mathbf{V}$, and the augmented response data are conditionally independently distributed.

The goal is to derive the conditional posterior density of $(\boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_{\boldsymbol{\theta}})$. It follows that the posterior distribution of $(\boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_{\boldsymbol{\theta}})$ is given by

$$\begin{aligned} L(\boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_{\theta}, \boldsymbol{\delta} \,|\, \boldsymbol{z}_{...}, \boldsymbol{y}_{...}, \boldsymbol{v}_{...}) &\propto p(\boldsymbol{z}_{...} \,|\, \boldsymbol{y}_{...}, \boldsymbol{v}_{...}, \boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_{\theta}, \boldsymbol{\delta}) p(\boldsymbol{v}_{...} \,|\, \boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{\delta}) \\ &= p(\boldsymbol{z}_{...} \,|\, \boldsymbol{y}_{...}, \boldsymbol{\theta}_{..}, \boldsymbol{\zeta}) p(\boldsymbol{v}_{...} \,|\, \boldsymbol{\delta}) \\ &\propto p(\boldsymbol{z}_{...} \,|\, \boldsymbol{y}_{...}, \boldsymbol{\theta}_{..}, \boldsymbol{\zeta}) \\ &\propto \prod_{k=1}^{K} \prod_{j=1}^{n_k} \prod_{i \in \mathcal{I}_{jk}} \exp\{-0.5(z_{ijk} - a_i \theta_{jk} + b_i)^2\} \\ &\quad \times \mathbb{1}_{(Z_{ijk}, y_{ijk})}, \end{aligned} \tag{6}$$

where $\boldsymbol{z}_{..} = (z_{11}, \ldots, z_{In})'$ and $\mathcal{I}_{jk}$ is the subset of items presented to (or answered by) respondent j from group k (they are related to the V_{ijk}, which are known). On the other hand, the joint prior distribution of the unknown parameters is assumed to be

$$p(\boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_{\theta} \,|\, \boldsymbol{\eta}_{\boldsymbol{\zeta}}, \boldsymbol{\eta}_{\boldsymbol{\eta}}) = \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} p(\theta_{jk} \,|\, \boldsymbol{\eta}_{\theta_k}) \right\} \left\{ \prod_{i=1}^{I} p(\boldsymbol{\zeta}_i \,|\, \boldsymbol{\eta}_{\boldsymbol{\zeta}}) \right\} \left\{ \prod_{k=1}^{K} p(\boldsymbol{\eta}_{\theta_k} \,|\, \boldsymbol{\eta}_{\boldsymbol{\eta}}) \right\},$$

where $\boldsymbol{\eta}_{\boldsymbol{\zeta}}$ and $\boldsymbol{\eta}_{\boldsymbol{\eta}}$ are the hyperparameters associated with $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}_{\theta}$, respectively. Moreover, we are assuming independence between items and population parameters. In the following, we will present two MCMC algorithms. The first uses the original density of the centered skew-normal distribution of the latent traits (Equation (2)), whereas the second considers a stochastic representation of that.

3.1 *An MCMC algorithm which uses the* $\mathbf{SN_C}$ *density*

The prior distribution of the latent traits is given by Equation (2). For the item parameters, a usual choice is [3]

$$p(\boldsymbol{\zeta}_i) \propto \exp[-0.5(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_{\zeta})^{\top} \boldsymbol{\Psi}_{\zeta}^{-1} (\boldsymbol{\zeta}_i - \boldsymbol{\mu}_{\zeta})] \mathbb{1}_{(a_i > 0)}. \tag{7}$$

For the transformed population parameters, we consider the following prior density:

$$p(\xi_{\theta_k}, \omega_{\theta_k}, \lambda_{\theta_k}) = p(\xi_{\theta_k}, \omega_{\theta_k}) p(\lambda_{\theta_k}) \propto \frac{1}{\omega_{\theta_k}} \left(1 + \frac{\lambda_{\theta_k}^2}{\mathrm{d}\varphi^2} \right)^{-(d+1)/2}, \tag{8}$$

where

$$\begin{aligned} p(\xi_{\theta_k}, \omega_{\theta_k}) &\propto \frac{1}{\omega_{\theta_k}}, \\ p(\lambda_{\theta_k}) &\propto \left(1 + \frac{\lambda_{\theta_k}^2}{d\varphi^2} \right)^{-(d+1)/2}. \end{aligned} \tag{9}$$

From Equation (8) we can obtain an approximation to the Jeffreys prior, considering $d = \frac{1}{2}$ and $\varphi^2 = (\pi^2/4)$, and the prior induced by $\delta_{\theta_k} \sim U(-1, 1)$ (making $d = 2$ and $\varphi^2 = \frac{1}{2}$). For simplicity, these two priors will be called Jeffreys prior and Uniform prior, respectively. Another

choice is based on the following joint prior density:

$$p(\xi_{\theta_k}, \omega_{\theta_k}, \lambda_{\theta_k}) = p(\xi_{\theta_k})p(\omega_{\theta_k})p(\lambda_{\theta_k}), \tag{10}$$

considering

$$\begin{aligned}
\xi_{\theta_k} &\sim N(\mu_{\xi_\theta}, \sigma^2_{\xi_\theta}),\\
\omega^2_{\theta_k} &\sim \text{IG}(\alpha_{\omega_\theta}, \beta_{\omega_\theta}),\\
\lambda_{\theta_k} &\sim N(\mu_{\lambda_\theta}, \sigma^2_{\lambda_\theta}),
\end{aligned}$$

where IG stands for the inverse-gamma distribution. Consequently, considering the augmented likelihood (6) and the priors (7) and (8), we have that the posterior distribution of $(\boldsymbol{Z}_{\cdots}, \boldsymbol{\theta}_{\cdot\cdot}, \boldsymbol{\zeta}, \boldsymbol{\eta}_\theta)$ is given by

$$\begin{aligned}
p(\boldsymbol{Z}_{\cdots}, \boldsymbol{\theta}_{\cdot\cdot}, \boldsymbol{\zeta}, \boldsymbol{\eta}_\theta \mid \boldsymbol{y}_{\cdots}, \boldsymbol{\eta}_\zeta, \boldsymbol{\eta}_{\boldsymbol{\eta}}) &\propto p(\boldsymbol{Z}_{\cdots} \mid \boldsymbol{\theta}_{\cdot\cdot}, \boldsymbol{\zeta}, \boldsymbol{y}_{\cdots})p(\boldsymbol{\theta}_{\cdot\cdot} \mid \boldsymbol{\eta}_\theta)p(\boldsymbol{\zeta} \mid \boldsymbol{\eta}_\zeta)p(\boldsymbol{\eta}_\theta \mid \boldsymbol{\eta}_{\boldsymbol{\eta}})\\
&= \left\{\prod_{k=1}^{K}\prod_{j=1}^{n_k}\prod_{i\in I_{jk}} p(z_{ijk} \mid \theta_{jk}, \boldsymbol{\zeta}_i, y_{ijk})\right\}\\
&\quad \times \left\{\prod_{k=1}^{K}\prod_{j=1}^{n_k} p(\theta_{jk} \mid \boldsymbol{\eta}_{\theta_k})\right\}\left\{\prod_{i=1}^{I} p(\boldsymbol{\zeta}_i)\right\}\\
&\quad \times \left\{\prod_{k=1}^{K} p(\xi_{\theta_k}, \omega_{\theta_k}, \lambda_{\theta_k})\right\}\\
&\propto \left\{\prod_{k=1}^{K}\prod_{j=1}^{n_k}\prod_{i\in I_{jk}} \exp\{-0.5(z_{ijk} - a_i\theta_{jk} + b_i)^2\}\, \mathbb{1}_{(Z_{ijk}, y_{ijk})}\right\}\\
&\quad \times \left\{\prod_{k=1}^{K} \omega_{\theta_k}^{-n_k}\prod_{j=1}^{n_k}\phi\left(\frac{\theta_{jk} - \xi_{\theta_k}}{\omega_{\theta_k}}\right)\Phi\left[\left(\frac{\theta_{jk} - \xi_{\theta_k}}{\omega_{\theta_k}}\right)\right]\right\}\\
&\quad \times \left\{\prod_{i=1}^{I} \exp\left[-0.5(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta)^\top \boldsymbol{\Psi}_\zeta{}^{-1}(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta)\right] \mathbb{1}_{(a_i>0)}\right\}\\
&\quad \times \left\{\prod_{k=1}^{K} \frac{1}{\omega_{\theta_k}}\left(1 + \frac{\lambda^2_{\theta_k}}{d\varphi^2}\right)^{-(d+1)/2}\right\}.
\end{aligned}$$

The posterior distribution, under the prior density (10), can be seen in [35]. Due to the augmented data scheme, the full conditional distributions of the item parameters and of the augmented data themselves are known and easy to sample from. However, for the latent traits and population parameters, the full conditionals are not simple to sample from. Therefore, we need to use auxiliary algorithms to sample from such distributions. We consider the *Metropolis–Hastings* within Gibbs sampling algorithm [27,28]. However, we are using the augmented data likelihood instead of the original likelihood. Then, we call our algorithm ADMHWGS (augmented data Metropolis–Hastings within Gibbs sampling), as in [4].

To implement the *Metropolis–Hastings* steps, we need to consider suitable proposal densities [17] for both latent traits and population parameters. For the latent traits, we consider

$$J_t(\theta_{jk}^{(*)} \mid \theta_{jk}^{(t-1)}) = N(\theta_{jk}^{(t-1)}, \sigma_\theta^2),$$

where $J_t(\cdot)$ stands for the proposal density at iteration t. For the population parameters, we consider

$$\begin{aligned} J_t(\xi_k^{(*)} \mid \xi_k^{(t-1)}) &= N(\xi_k^{(t-1)}, \sigma_\xi^2), \\ J_t(\omega_k^{(*)} \mid \omega_k^{(t-1)}) &= \text{Log-Normal}(\omega_k^{(t-1)}, \sigma_\omega^2), \\ J_t(\lambda_k^{(*)} \mid \lambda_k^{(t-1)}) &= N(\lambda_k^{(t-1)}, \sigma_\lambda^2), \end{aligned}$$

where $(\sigma_\theta^2, \sigma_\xi^2, \sigma_\omega^2, \sigma_\lambda^2)$ are fixed in advance. Denoting $(\cdot)$ the set of all other parameters, the proposed algorithm (ADMHWGS), for $t = 1, 2, \ldots, B, \ldots, M$, where B is the burn-in and M is the generated sample size, simulates iteratively all unknown quantities in the following order:

- Start the algorithm by choosing suitable starting values.
- Simulate Z_{ijk} from $Z_{ijk} \mid (\cdot)$, $k = 1, \ldots, K$, $j = 1, \ldots, n_k$ and $i \in \mathcal{I}_{jk}$.
- Simulate θ_{jk} from $\theta_{jk} \mid (\cdot)$, $k = 1, \ldots, K$ and $j = 1, \ldots, n_k$.
- Simulate $\boldsymbol{\zeta}_i$ from $\boldsymbol{\zeta}_i \mid (\cdot)$, $i \in \mathcal{I}_{jk}$.
- Simulate ξ_{θ_k} from $\xi_{\theta_k} \mid (\cdot)$, $k = 1, \ldots, K$.
- Simulate ω_{θ_k} from $\omega_{\theta_k} \mid (\cdot)$, $k = 1, \ldots, K$.
- Simulate λ_{θ_k} from $\lambda_{\theta_k} \mid (\cdot)$, $k = 1, \ldots, K$.

The population parameters of interest $(\mu_{\theta_k}, \sigma_{\theta_k}^2, \delta_{\theta_k})$ can be easily recovered by considering the relationships given by Equation (3). In the following, we will present an MCMC algorithm based on a hierarchical representation for the SN_C distribution.

3.2 *An MCMC algorithm with a hierarchical structure for the* SN_C *distribution*

One can consider the usual hierarchical representation of the centered skew-normal distribution [4], that is,

$$\theta_{jk} = \xi_{\theta_k} + \omega_{\theta_k} \left(\delta_{\theta_k} H_{jk} + \sqrt{1 - \delta_{\theta_k}^2} Q_{jk} \right), \tag{11}$$

where $H_{jk} \sim \text{HN}(0, 1)$, $Q_{jk} \sim N(0, 1)$, $H_{jk} \perp Q_{jk}$, $\forall j, k$, and $\text{HN}(0, 1)$ denotes a half-normal distribution based on an $N(0, 1)$ distribution. We can also rewrite Equation (11) as

$$\theta_{jk} \mid (h_{jk}, \boldsymbol{\eta}_{\theta_k}) \sim N(\xi_{\theta_k} + \tau_{\theta_k} h_{jk}, \varsigma_{\theta_k}^2), \tag{12}$$

$$H_{jk} \sim \text{HN}(0, 1), \tag{13}$$

where

$$\tau_{\theta_k} = \omega_{\theta_k} \delta_{\theta_k}, \tag{14}$$

$$\varsigma_{\theta_k}^2 = \omega_{\theta_k}^2 (1 - \delta_{\theta_k}^2). \tag{15}$$

Considering now the population parameters, we can note that the prior defined by Equation (9) can be rewritten as follows:

$$\begin{aligned} \lambda_{\theta_k} \mid t_k &\sim N\left(0, \frac{\varphi^2}{t_k}\right), \\ T_k &\sim \text{gamma}\left(\frac{d}{2}, \frac{d}{2}\right). \end{aligned}$$

Thus, the prior showed in Equation (8) can be rewritten as

$$p(\lambda_{\theta_k}, \xi_{\theta_k}, \omega_{\theta_k}, t_k) \propto \frac{1}{\omega_{\theta_k}} \exp\left(-\frac{1}{2}\frac{\lambda_{\theta_k}^2 t_k}{\varphi^2}\right) t_k^{((d+1)/2)-1} \exp\left(-\frac{d}{2}t_k\right).$$

Considering the aforementioned reparametrization, we have

$$p(\xi_{\theta_k}, \tau_{\theta_k}, \varsigma_{\theta_k}, t_k) \propto \frac{1}{\omega_{\theta_k}} \exp\left(-\frac{1}{2}\frac{\tau_{\theta_k}^2 t_k}{\varsigma_{\theta_k}^2 \varphi^2}\right) t_k^{((d+1)/2)-1} \exp\left(-\frac{d}{2}t_k\right). \tag{16}$$

Equation (16) is a joint prior for the parameters $(\xi_{\theta_k}, \tau_{\theta_k}, \varsigma_{\theta_k})$ and the latent variables T_k. On the other hand, we may note that Equation (12) can be viewed as a regression model with the response variable θ_{jk}, intercept ξ_{θ_k} and slope τ_{θ_k}. Therefore, analogously to what was made for the item parameters, we can consider a bivariate normal distribution as a prior distribution for the vector $(\xi_{\theta_k}, \tau_{\theta_k})^\top$, which we will denote by $\boldsymbol{\beta}_{\theta_k}$, and an inverse gamma prior for parameter $\varsigma_{\theta_k}^2$. That is,

$$\boldsymbol{\beta}_{\theta_k} \sim N(\boldsymbol{\mu}_\theta, \Sigma_\theta), \tag{17}$$

$$\varsigma_{\theta_k}^2 \sim \text{IG}(\alpha_{\varsigma_\theta}, \beta_{\varsigma_\theta}), \tag{18}$$

where

$$\boldsymbol{\beta}_{\theta_k} = \begin{pmatrix} \xi_{\theta_k} \\ \tau_{\theta_k} \end{pmatrix}, \quad \boldsymbol{\mu}_\theta = \begin{pmatrix} \mu_{\xi_\theta} \\ \mu_{\tau_\theta} \end{pmatrix} \quad \text{and} \quad \Sigma_\theta = \begin{pmatrix} \sigma_{\xi_\theta}^2 & \rho\sigma_{\xi_\theta}^2\sigma_{\tau_\theta}^2 \\ \rho\sigma_{\xi_\theta}^2\sigma_{\tau_\theta}^2 & \sigma_{\tau_\theta}^2 \end{pmatrix}.$$

We will assume the following structure for the joint prior distribution of the parameters:

$$\begin{aligned} p(\boldsymbol{\theta}_{\cdot\cdot}, \boldsymbol{h}_{\cdot\cdot}, \boldsymbol{\zeta}, \boldsymbol{\eta}_\theta, \boldsymbol{t}_\cdot \mid \boldsymbol{\eta}_\zeta, \boldsymbol{\eta}_\eta) = & \left\{\prod_{k=1}^{K}\prod_{j=1}^{n_k} p(\theta_{jk} \mid \boldsymbol{\eta}_{\theta_k})\right\}\left\{\prod_{k=1}^{K}\prod_{j=1}^{n_k} p(h_{jk})\right\} \\ & \times \left\{\prod_{i=1}^{I} p(\boldsymbol{\zeta}_i \mid \boldsymbol{\eta}_\zeta)\right\}\left\{\prod_{k=1}^{K} p(\boldsymbol{\eta}_{\theta_k} \mid \boldsymbol{\eta}_\eta)\right\} \\ & \times \left\{\prod_{k=1}^{K} p(t_k)\right\}, \end{aligned} \tag{19}$$

where $\boldsymbol{h}_{\cdot\cdot} = (\boldsymbol{h}_{\cdot 1}, \ldots, \boldsymbol{h}_{\cdot K})^\top$, $\boldsymbol{h}_{\cdot k} = (h_{1k}, \ldots, h_{n_k k}), k = 1, 2, \ldots, K$ and $\boldsymbol{t}_\cdot = (t_1, \ldots, t_K)^\top$. The prior distributions for the latent traits and the variables $\boldsymbol{h}$ are given by Equations (12) and (13), respectively. For the item parameters, we will consider the prior given by Equation (7) as before. For the population parameters, we can consider either the joint prior defined by Equation (16) or the prior distributions presented in Equations (17) and (18). Depending on the prior distribution adopted, the posterior distribution will be different. They can be see in [35]. It is important to observe that the hierarchical representation makes it possible to obtain the full conditional distributions with known forms to all parameters. Therefore, it is possible to use the full Gibbs sampling algorithm. This algorithm will be called augmented data Gibbs sampling (ADGS).

For each one of the two posterior distributions, respectively, we will have the following algorithm if we consider the posterior distribution (A1):

- Start the algorithm by choosing suitable initial values.
- Simulate Z_{ijk} from $Z_{ijk} \mid (\cdot)$, $k = 1, \ldots, K, j = 1, \ldots, n_k$ and $i \in I_{jk}$.

- Simulate h_{jk} from $H_{jk} \mid (\cdot)$, $k = 1, \ldots, K$ and $j = 1, \ldots, n_k$.
- Simulate θ_{jk} from $\theta_{jk} \mid (\cdot)$, $k = 1, \ldots, K$ and $j = 1, \ldots, n_k$.
- Simulate $\boldsymbol{\zeta}_i$ from $\boldsymbol{\zeta}_i \mid (\cdot)$, $i \in I_{jk}$.
- Simulate t_k from $T_k \mid (\cdot)$, $k = 1, \ldots, K$.
- Simulate ξ_{θ_k} from $\xi_{\theta_k} \mid (\cdot)$, $k = 1, \ldots, K$.
- Simulate τ_{θ_k} from $\tau_{\theta_k} \mid (\cdot)$, $k = 1, \ldots, K$.
- Simulate $\varsigma^2_{\theta_k}$ from $\varsigma^2_{\theta_k} \mid (\cdot)$, $k = 1, \ldots, K$.

On the other hand, if we consider the posterior distribution (A2), we have the following algorithm:

- Start the algorithm by choosing suitable initial values.
- Simulate Z_{ijk} from $Z_{ijk} \mid (\cdot)$, $k = 1, \ldots, K$, $j = 1, \ldots, n_k$ and $i \in I_{jk}$.
- Simulate h_{jk} from $H_{jk} \mid (\cdot)$, $k = 1, \ldots, K$ and $j = 1, \ldots, n_k$.
- Simulate θ_{jk} from $\theta_{jk} \mid (\cdot)$, $k = 1, \ldots, K$ and $j = 1, \ldots, n_k$.
- Simulate $\boldsymbol{\zeta}_i$ from $\boldsymbol{\zeta}_i \mid (\cdot)$, $i \in I_{jk}$.
- Simulate $\boldsymbol{\beta}_{\theta_k}$ from $\boldsymbol{\beta}_{\theta_k} \mid (\cdot)$, $k = 1, \ldots, K$.
- Simulate $\varsigma^2_{\theta_k}$ from $\varsigma^2_{\theta_k} \mid (\cdot)$, $k = 1, \ldots, K$.

For more details, see the appendix and [35]. Again, the population parameters of interest can be easily recovered by using the relationships given by Equations (3), (14) and (15). All algorithms were implemented in a program developed by the authors using the object-oriented statistical system Ox^{tm} (see http://www.doornik.com/products.html#Ox), and it is available upon request from the authors.

4. Simulation study

In this section, we conducted two simulation studies. First, we discuss convergence aspects of the two algorithms previously presented, comparing them under simulate samples. Second, we performed a replication study with the selected algorithm (that one which present the best convergence proprieties) in order to evaluate it in terms of parameter recovery.

4.1 *Convergence study*

The following structure for the tests was considered (in order to have a similar situation concerning the real data set analyzed, as we will show ahead, including the values for the item and population parameters):

- test 1: 20 items;
- test 2: test 1 + 20 other items;
- test 3: the last 20 items of test 2 + 20 other items;
- test 4: the last 20 items of test 3 + 20 other items.

This linking design (common items among the tests) is necessary to ensure the model identifiability and then to allow the results from the different groups to lie on the same scale (see Section 2.1). The item parameters were fixed in the following intervals: $a_i \in [0.7, 1.4]$ and $b_i^* \in [-2, 4]$. The latent traits were simulated from independent skew-normal distributions (each one representing one group) under CP, with means $\boldsymbol{\mu}_\theta = (0.0, 1.0, 1.4, 2.0)^\top$, standard deviations $\boldsymbol{\sigma}_\theta = (1.00, 0.88, 0.62, 0.77)^\top$ and asymmetry coefficients $\boldsymbol{\gamma}_\theta = (0.00, 0.14, 0.50, -0.50)^\top$ for groups 1–4, respectively. Therefore, we have unidimensional tests with common items. Note that the number of subjects per group were $(n_1 = 556, n_2 = 556, n_3 = 401, n_4 = 294)$. We tried

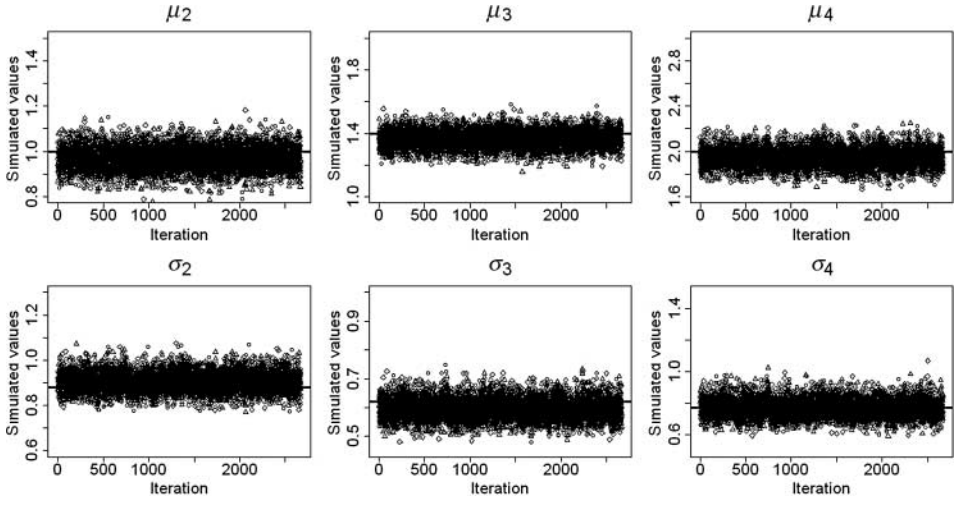

Figure 2. Trace plots for the population parameters: ADGS algorithm. ∘, set 1; △, set 2; ⋄, set 3; —, true value.

to mimic the results obtained by Azevedo *et al.* [3], given that we will analyze the same data set. The values of the asymmetry parameter were chosen in order to consider symmetry, positive weak asymmetry, positive strong asymmetry and negative right strong asymmetry for the latent trait distributions.

For the prior of the item parameters, we considered $(1, 0)$ as a mean vector and $\text{diag}(0.5, 9)$ as the covariance matrix. For the population parameters, we considered the Jeffreys prior. For this simulated data set, we generated three chains, based on different sets of starting values, for each one of the two algorithms (ADMHWGS and ADGS). The Gelman–Rubin [18] statistics ranged from 1.00 to 1.02 for both algorithms. The closer to 1 the observed values of this statistic, the stronger is the evidence toward the convergence of the chains. Also, an inspection of trace plots indicated that the chains mixed very well for both algorithms (Figure 2; for the ADGS algorithm). Concerning the ADMHWGS, we did not present the traceplots, but the conclusions are the same. Further, the correlograms (not shown) revealed that the samples produced by storing values at every 30th iteration are enough to produce samples with negligible autocorrelations.

Furthermore, we compared the two algorithms concerning the effective sample size (ESS) [23,31]. Such statistic is defined for each parameter as the number of the MCMC samples drawn, M, divided by the parameter's autocorrelation time, $\gamma = 1 + 2\sum_{b=1}^{\infty} \rho_b$, where ρ_b is the autocorrelation at the time b. In order to estimate γ we consider $(1 + \rho^*)/(1 - \rho^*)$, where $\rho^* = \max_{b \geq 1} |\rho_b|$, as in [31]. Table 1 presents the values of the ESS and ESS per minute (ESS/m) considering four different prior distributions (see Section 4.2 for more details concerning these priors). We can see that, in a general way, the AGDS algorithm presented better results than the ADMHWGS algorithm, since the higher the EES, the better is the performance of the algorithm.

4.2 *Replication study (parameter recovery)*

Due to the results of Section 4.1, we selected the ADGS algorithm to perform the parameter recovery study. We generated other $R = 10$ replicas (data sets), for each one of the situation defined by crossing the levels of different factors. These factors were (with the levels within parenthesis): number of respondents per group (NE) (1500, 3000), number of items per group (NI) (20, 40), number of common items (NCI) (25%, 50%) and prior distributions (P) (prior1,

Table 1. ESS and ESS per minute for the two algorithms.

	ADMHWGS		ADGS		
Prior	ESS	ESS/m	ESS	ESS/m	Ratio
1	4474.178	25.228	4771.270	36.361	1.441
2	4549.746	20.190	4724.984	36.014	1.784
3	5054.523	32.316	4955.711	35.533	1.010
4	4569.048	27.968	5103.164	23.491	0.840

Table 2. Hyperparameters chosen for priors defined by Equations (7), (8), (17) and (18).

	Hyperparameters					
Prior	$\boldsymbol{\mu}_\zeta$	$\boldsymbol{\Psi}_\zeta$	$(d;\varphi^2)$	$(\mu_{\xi_\theta},\sigma^2_{\xi_\theta})$	$(\mu_{\tau_\theta},\sigma^2_{\tau_\theta})$	$(\alpha_{\varsigma^2_\theta},\beta_{\varsigma^2_\theta})$
1	$(1, 0)$	$(0.5, 9)$	$(0.5, \pi^2/4)$	–	–	–
2	$(1, 0)$	$(0.5, 9)$	$(2, 0.5)$	–	–	–
3	$(1, 0)$	$(0.5, 9)$	–	$(0, 0.7^2)$	$(0, 0.8^2)$	$(3.4, 3)$
4	$(1, 0)$	$(0.5, 9)$	–	$(0, 3.1^2)$	$(0, 0.23^2)$	$(2.1, 2.32)$

prior2, prior3, prior4). For the NCI factor, the levels correspond to the percentage of the total number of items in each test which are common between two tests. The priors from 1 to 4 are, in fact, sets of priors. For simplicity, these will be named priors only (without mentioning that they are sets of priors explicitly). Such sets differ only in terms of the priors for the population parameters as described in Table 2. While priors 1 and 2 are Jeffrey priors themselves, the priors 3 and 4 are conditional conjugate priors.

In addition, we defined another factor called asymmetry of the reference group (ARG) with levels 1–3. These levels correspond to situations where the ARG distribution is null, strongly positive and strongly negative, respectively. As explained before, the reference group is group 1. The other population parameters were defined as follows: the means and the standard deviations were fixed at $\mu_\theta = (0, -1, 1)$ and $\sigma_{\theta_k} = (1, 0.8, 1.2)$ for the groups 1–3, respectively. The asymmetry coefficients for the groups 2 and 3 were fixed at $\gamma_{\theta_2} = 0.6$ and $\gamma_{\theta_3} = -0.6$. That is, we are considering situations where the latent traits distribution of the reference group is symmetric or asymmetric and for the other groups the distributions are asymmetric. Table 3 presents the values for the population parameters for each group in each level of the ARG factor.

The values of the item parameters were chosen considering the values of the population parameters (concerning the difficulty parameter) and in order to have items with different discrimination powers and different difficulty levels. We decided to consider only three groups (instead of four groups as before) due to the time necessary to run all the replicas. We considered the RMSE (root-mean-square error) and the standard error (SE) of the Monte Carlo replicas to measure the accuracy of the estimates, based on the estimates obtained with the 10 replicas. These two statistics are defined by

$$\mathrm{SE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\vartheta}_r - \bar{\hat{\vartheta}})^2}; \quad \mathrm{RMSE} = \sqrt{\mathrm{SE}^2 + (\bar{\hat{\vartheta}} - \vartheta)^2},$$

where $\vartheta = (\theta, \zeta, \eta)$, $\hat{\vartheta}_r$ is the estimate (posterior expectation) obtained in the replica r, $\bar{\hat{\vartheta}} = (1/R)\sum_{r=1}^{R}\hat{\vartheta}_r$, $R = 10$ and ϑ is the true value.

Table 3. Population parameters chosen for each group and each level of the ARG factor.

		ARG = 1	ARG = 2	ARG = 3
Group 1	μ_{θ_1}	0	0	0
	σ_{θ_1}	1	1	1
	γ_{θ_1}	0	0.6	−0.6
Group 2	μ_{θ_2}	−1	−1	−1
	σ_{θ_2}	0.8	0.8	0.8
	γ_{θ_2}	0.6	0.6	0.6
Group 3	μ_{θ_3}	1	1	1
	σ_{θ_3}	1.2	1.2	1.2
	γ_{θ_3}	−0.6	−0.6	−0.6

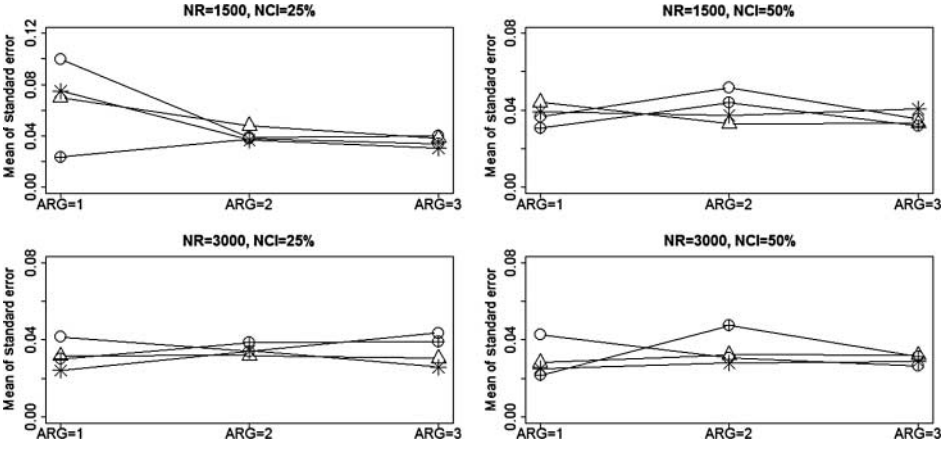

Figure 3. Mean of standard error for the population asymmetry coefficients. –○–, Prior 1; –△–, Prior 2; –∗–, Prior 3; –⊕–, Prior 4.

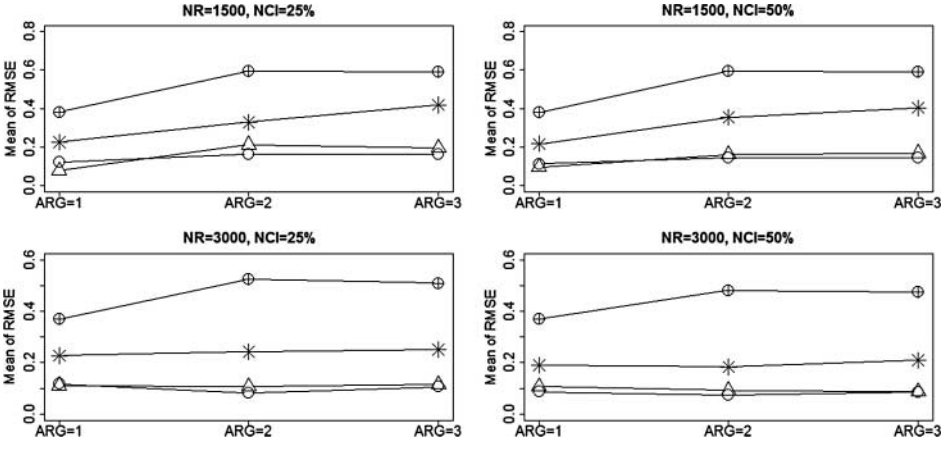

Figure 4. Mean of RMSE for the population asymmetry coefficients. –○–, Prior 1; –△–, Prior 2; –∗–, Prior 3; –⊕–, Prior 4.

From an inspection of Figures 3 and 4, we can conclude that the ADGS algorithm recovered all parameters properly. In addition, one can see that there is a slight difference between the results obtained from the different priors. This difference is much higher for the asymmetry coefficient and, in this case, the more accurate results were obtained by using the Jeffreys prior. Moreover, there is strong indication that the higher the NCI values (simultaneously), the more accurate are the results. The results related to the other parameters (not showed), that is, latent traits, item parameters and the other population parameters, confirm the conclusions we presented.

5. Real data analysis

The description of the data set analyzed was adapted from [3]. The International Project on Mathematical Attainment (IPMA) is a major longitudinal educational assessment coordinated by the University of Exeter, England, through the Centre for Innovation in Mathematics Teaching (CMIT) toward Mathematics achievement. Several countries (including Brazil) are participating in this study, through the State University of Londrina. The IPMA monitors the mathematical progress of pupils in primary schools and relates the progress to several factors including style of teaching and curriculum organization. The aim is to provide recommendations for good practice in primary mathematics education. The math tests are designed to assess progress on key mathematical topics and concepts. The content and the difficulty of the items are equivalent to primary school level.

In Brazil, 568 first-grade students were selected from eight public primary schools. The eight schools were chosen from different places of Londrina city. There were six municipal schools and two state schools. The number of selected students per school varied and students belonged to different classes. The students are nested in classes and classes are nested in schools. Several student-level and teacher-level background variables (such as gender, skin color, age, education level of the parents) were collected. Even though it is expected to observe some relationship among the latent traits and the background variables as well as among the latent traits of the subjects belonging to the same classes, we will not incorporate this information, as in [16]. This deserves more investigation, but is far beyond the scope of this paper, by developing a skew multiple group multilevel modeling.

The first group comprised 568 students but, along the subsequent grades, some students dropped out from the study for different reasons. The present data set consists of the following number of students, from the first up to the fourth grade: 556, 556, 401 and 295. The students from the first grade answered a test of 20 items, the second graders of 40 items, including the 20 items of the first-grade test. The third and fourth graders responded to a test of 60 items, including the 40 items of the second-grade test, and 80 items, including the third-grade items, respectively. All items were corrected as right–wrong items.

For grades 2 to 4, the responses to the 20 new items and the preceding 20 test items are considered, which leads to 40 items for each grade and a total of 60 test items. For grade 1, the responses to the 20 items are considered. This way a slightly more balanced test design is constructed, which will lead to more comparable standard errors of the latent math scores across grades. In conclusion, we have four groups (corresponding to each one of the grades) and a total of 80 different items. The data set is available upon request from authors. Further details can be obtained in [30]. The whole data set was previously analyzed by Azevedo *et al.* [3] using the symmetric normal MGM. We compared those results with the ones obtained by using the model proposed in this paper, estimated by using the ADGS algorithm with the Jeffreys prior. We considered the model fit assessment tools described in [3], and for model comparison, we used the statistics considered in [4].

Even though we are dealing with a longitudinal design, the results obtained by Azevedo *et al.* [3] revealed that the within-subject correlations of the latent traits are not significant. Therefore, an

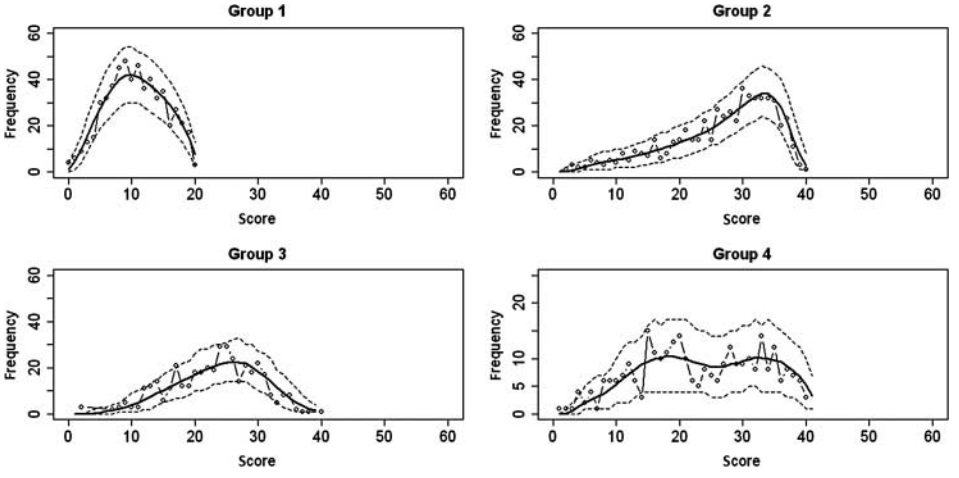

Figure 5. Observed and predictive scores distribution (skew model). Solid line, predicted scores; –○–, observed scores; – – –, 95% credibility interval.

MGM (i.e. an independent model) can be considered. Also, we did not perform any dimensionality test since we have a linking design that can be recognized as an incomplete block design, which makes more difficult the determination of the dimensionality of the tests. This point deserves further investigation but it is beyond the scope of this work.

Following [3], we considered the Bayesian p-value based on the deviance residual as a global measure of goodness of fit. The result was $p = 0.7910$, which indicates that it cannot be concluded that the SMGIRT model does not fit the data. Figure 5 presents the observed and predicted scores with 95% Bayesian credibility intervals for the four groups. It can be seen that all the observed

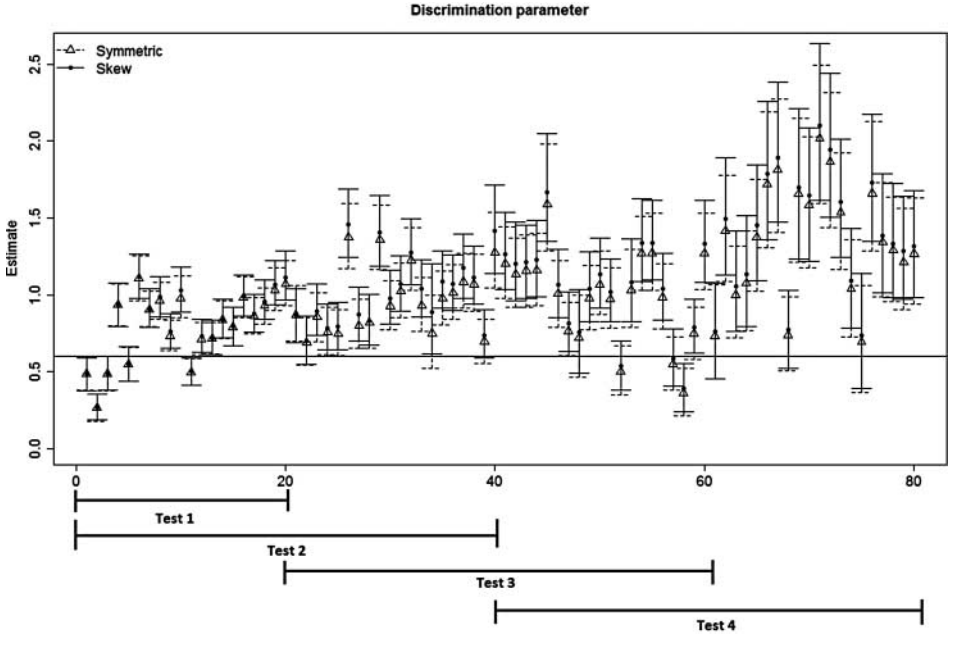

Figure 6. Discrimination parameter estimates and 95% credibility intervals (skew and symmetric models).

score distributions are well within the intervals. Therefore, we can conclude that the model is well fitted to the data. The skew-normal assumption of the latent traits distributions was verified using the goodness-of-fit test proposed by Cabras and Castellanos [13]. In this test, the Kolmogorov–Smirnov's statistic is utilized for measuring the difference between an empirical distribution and a theoretical distribution, which must be some element of the skew-normal class. The *p*-values obtained were greater than 0.100 for groups 1–3 and between 0.010 and 0.025 for group 4, indicating that only the latent trait distribution of group 4 is not compatible with the skew-normal model.

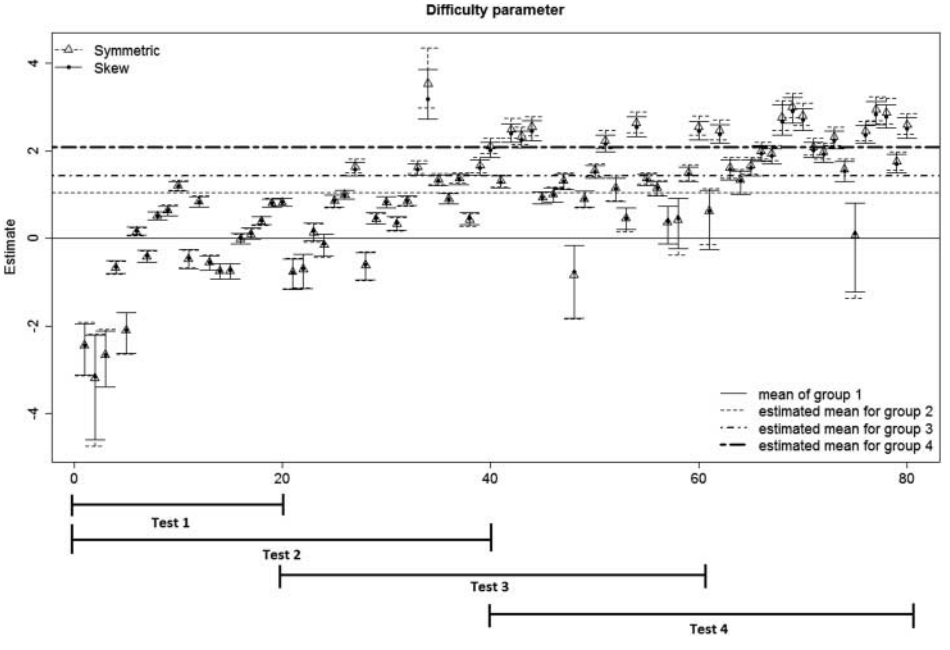

Figure 7. Difficulty parameter estimates and 95% credibility intervals (skew and symmetric models).

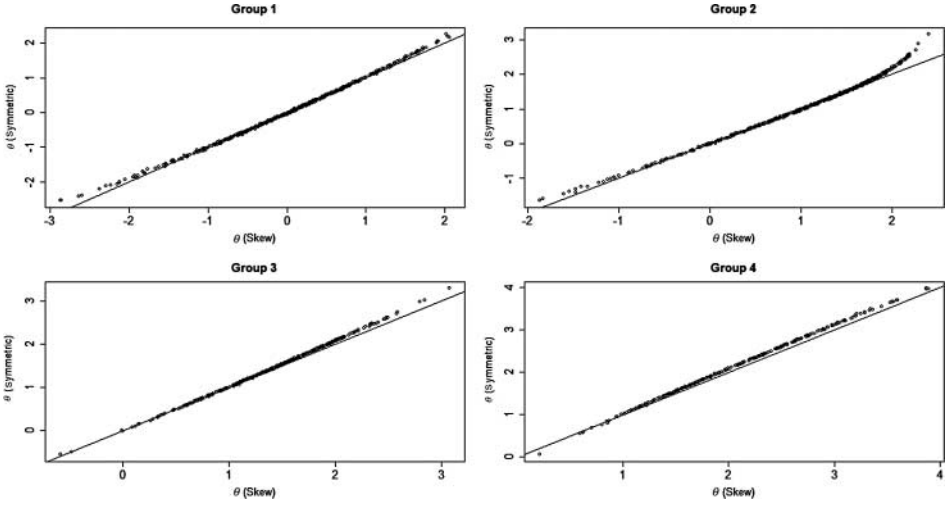

Figure 8. Scatter plot for the latent trait estimates (skew and symmetric models).

Table 4. Estimates and credibility intervals for the population parameters (skew and symmetric models).

	Skew						Symmetric					
Parameter	EAP	PSD	CI (95%)		HPD (95%)		EAP	PSD	CI (95%)		HPD (95%)	
μ_{θ_1}	0	–	–	–	–	–	0	–	–	–	–	–
μ_{θ_2}	1.049	0.050	0.956	1.150	0.958	1.153	1.078	0.054	0.974	1.186	0.981	1.193
μ_{θ_3}	1.427	0.057	1.317	1.539	1.316	1.536	1.463	0.060	1.356	1.588	1.356	1.589
μ_{θ_4}	2.093	0.088	1.926	2.270	1.917	2.258	2.156	0.091	1.986	2.346	1.996	2.354
σ_{θ_1}	1	–	–	–	–	–	1	–	–	–	–	–
σ_{θ_2}	0.828	0.045	0.740	0.917	0.750	0.923	0.887	0.042	0.811	0.970	0.810	0.967
σ_{θ_3}	0.588	0.036	0.520	0.659	0.519	0.659	0.623	0.037	0.556	0.703	0.555	0.702
σ_{θ_4}	0.744	0.055	0.642	0.869	0.636	0.853	0.772	0.056	0.673	0.892	0.679	0.895
γ_{θ_1}	−0.283	0.153	−0.587	−0.001	−0.552	0.001	0.000	–	–	–	–	–
γ_{θ_2}	−0.805	0.092	−0.960	−0.605	−0.975	−0.630	0.000	–	–	–	–	–
γ_{θ_3}	−0.160	0.160	−0.484	0.045	−0.448	0.060	0.000	–	–	–	–	–
γ_{θ_4}	0.140	0.184	−0.079	0.569	−0.098	0.537	0.000	–	–	–	–	–

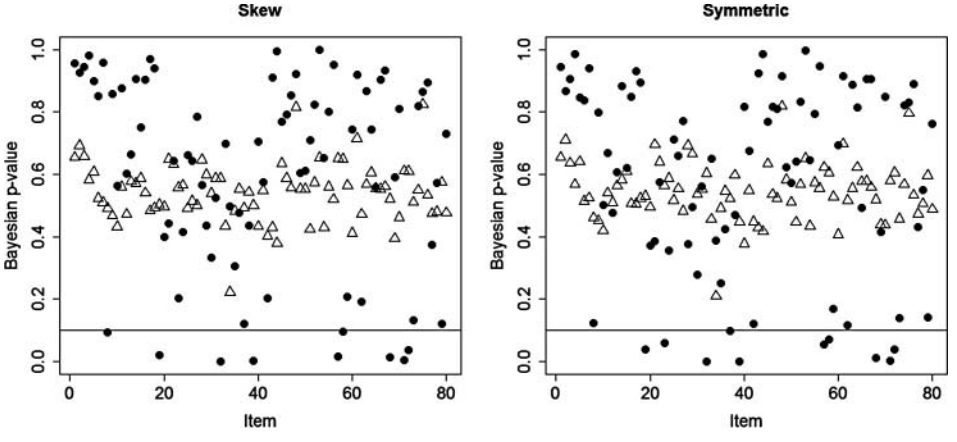

Figure 9. Bayesian *p*-values for the discrepancy measure based on the Pearson and deviance residuals. •, Pearson; △, deviance.

From Figures 5–8 and Table 4, we can see that, in general, the results from the two models were similar. However, for the second group (group 2), which presented the highest absolute value for the asymmetry coefficient (−0.805), the results were significantly different. For example, in Figure 8, we can see that there is a considerable difference between the estimates of the latent traits of group 2. The item fit was evaluated by using Bayesian *p*-values for both Pearson and deviance residual discrepancy measures, as in [3]. Items with *p*-values below 0.10 were considered not well fitted by the model. Figure 9 indicates that the two models fitted properly most of the items, even though more items were well fitted by considering the skew model than the symmetric model. In addition, the skew model tends to present higher estimates for the discrimination parameter compared with the symmetric model, as shown Figure 6. That is, under the skew model, the items present a higher discrimination power than in the symmetric model, which is an important feature. Finally, we compared the two models using the deviance information criteria (DIC), the expectation values of the Akaike's information criteria (EAIC) and Bayesian information criteria (EBIC) [37]. According to Table 5, all statistics indicated that our model fitted better the data set. Therefore, inferences based on our model are more reliable than those obtained by using the symmetric normal model.

Table 5. Statistics for model comparisons.

Model	$\widehat{D(\bar{\vartheta})}$	$\widehat{D(\vartheta)}$	$\hat{\rho}_D$	$\widehat{\text{DIC}}$	$\widehat{\text{EAIC}}$	$\widehat{\text{EBIC}}$
Symmetric	58103.00	60073.00	1969.80	62042.60	64012.60	83477.08
Skew	57873.00	59836.00	1962.60	61798.20	63761.20	83154.54

In a general way, the latent traits, the item parameters and the population parameters have the same interpretations in both models (skew and symmetric ones), except for the asymmetry coefficient. By using the skew model, we obtain more reliable estimates and a better characterization of the latent trait distributions. From the results of the skew model, we can see an increase in the mean of the latent traits (which indicates that the subjects present a gain in their knowledge). Also, we can note that the variance decreases up to the third grade and then increases again. However, in the last three grades, the subjects are more homogenous compared with the first grade. Furthermore, the latent trait distributions are symmetric in the first and third grades, negative asymmetric in the second grade and then positive asymmetric in the four grade. This indicates that the knowledge of the subjects tends to concentrate either around the mean or below it. All tests, in general, presents a reasonable discrimination power ($a > 0.6$) and get more difficulty, since the values of the difficulty parameter tend to be higher than the mean of the latent traits.

6. Final conclusions and remarks

We presented a multiple group IRT model with a centered skew-normal structure for the latent trait distributions. Such approach is more flexible than the usual standard normal one and more straightforwardly interpretable than the nonparametric ones (in terms of the density obtained at the final step of the estimation process). Moreover, it leads to an identified model, at least in the case of dichotomous IRT models. We developed two MCMC algorithms for the model fit and compared them concerning convergence issues. The selected algorithm (AGDS, which is a full Gibbs sampling algorithm) showed to be efficient in terms of parameter recovery, according to the simulation study. Furthermore, it was shown that not considering the asymmetry behavior of the latent trait distributions can lead to misleading estimates, depending on the level of asymmetry. In addition, the results indicated that the proposed model presented a better fit than the usual two-parameter model, for the real data analysis. Moreover, our approach indicated that three of the four groups present negative asymmetric behavior. In conclusion, our approach showed to be a promising alternative to the usual ones in analyzing multiple group IRT data sets. For future research we intend to explore some extensions of our model in order to consider other IRFs to analyze longitudinal data sets and results of multidimensional tests. Further investigation is also needed to consider alternative estimation algorithms for parameter estimation.

Acknowledgements

The authors thank two anonymous referees for their insightful comments and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the financial support.

References

[1] J. Albert, *Bayesian estimation of normal ogive item response curves using Gibbs sampling*, J. Educ. Behav. Stat. 17 (1992), pp. 251–269.

[2] C.L.N. Azevedo and D.F. Andrade, *An estimation method for latent traits and population paremeter for the nominal response model*, Braz. J. Probab. Stat. 24 (2010), pp. 415–433.

[3] C.L.N. Azevedo, D.F. Andrade and J.-P. Fox, *A Bayesian generalized multiple group IRT model with model-fit assessment tools*, Comput. Stat. Data Anal. 56 (2012), pp. 4399–4412.

[4] C.L.N. Azevedo, H. Bolfarine, and D.F. Andrade, *Bayesian inference for a skew-normal IRT model under the centred parameterization*, Comput. Stat. Data Anal. 55 (2011), pp. 353–365.

[5] C.L.N. Azevedo, H. Bolfarine, and D.F. Andrade, *Parameter recovery for a skew-normal IRT model under a Bayesian approach: Hierarchical frame-work, prior and kernel sensitivity and sample size*, J. Statist. Comput. Simul. 82 (2012), pp. 1679–1699.

[6] C.L.N. Azevedo, H. Bolfarine, and D.F. Andrade, *A note on identifiability and metric issues for skew IRT models*, Tech. Rep. rp02-12, Department of Statistics, University of Campinas, Campinas, 2012.

[7] C.L.N. Azevedo, J.-P. Fox, and D.F. Andrade, *Bayesian latent variable modeling of longitudinal item response data*, Stat. Comput., under review.

[8] C.L.N. Azevedo and H. Migon, *Bayesian inference in an item response theory model with a generalized student t link function*, XI Brazilian Meeting on Bayesian Statistics: EBEB 2012, Amparo-SP, Brazil, 2012.

[9] A. Azzalini, *A class of distribution which includes the normal ones*, Scand. J. Stat. 12 (1985), pp. 171–178.

[10] F.B. Baker and S.-H. Kim, *Item Response Theory: Parameter Estimation Techniques*, 2nd ed., Marcel Dekker, Inc., New York, 2004.

[11] J.L. Bazán, M.D. Branco, and H. Bolfarine, *A skew item response model*, Bayesian Anal. 1 (2006), pp. 861–892.

[12] D.R. Bock and M.F. Zimowski, *The multiple groups IRT*, in *Handbook of Modern Item Response Theory*, W.J. van der Linden and R.K. Hambleton, eds., Springer-Verlag, New York, 1997, pp. 433–448.

[13] S. Cabras and M. Castellanos, *Default Bayesian goodness-of-fit tests for the skew-normal model*, J. Appl. Stat. 36 (2009), pp. 223–232.

[14] R.J. De Ayala and M. Sava-Bolesta, *Item parameter recovery for the nominal response model*, Appl. Psychol. Meas. 23 (1999), pp. 3–19.

[15] C. E. DeMars, *Sample size and the recovery of nominal response model item parameters*, Appl. Psychol. Meas. 27 (2003), pp. 275–288.

[16] J.-P. Fox and C.A.W. Glas, *Bayesian estimation of a multilevel IRT model using Gibbs sampling*, Psychometrika 66 (2001), 271–288.

[17] D. Gamerman and H.F. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd ed., Chapman & Hall, London, 2006.

[18] A. Gelman and D.B. Rubin, *Inference from iterative simulation using multiple sequences*, Statist. Sci. 7 (1992), pp. 457–472.

[19] J.A. Gilford and H. Swaminathan, *Bias and the effect of priors in Bayesian estimation of item response models*, Appl. Psychol. Meas. 14(1) (1990), pp. 33–43.

[20] M.R. Harwell and J.E. Janosky, *An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in Bilog*, Appl. Psychol. Meas. 15 (1991), pp. 279–291.

[21] N. Henze, *A probabilistic representation of the skew-normal distribution*, Scand. J. Stat. 13 (1986), pp. 271–275.

[22] S.R. Hesketh, A. Skrondal, and A. Pickles, *Generalized multilevel structural equation modeling*, Psychometrika 69 (2004), pp. 167–190.

[23] R.E. Kass, B.P. Carlin, A. Gelman, and R. Neal, *Markov chain Monte Carlo in practice: A roundtable discussion*, Amer. Statist. 52 (1998), pp. 93–100.

[24] L. Kirisci, T.-C. Hsu, and L. Yu, *Robustness of item parameter estimation programs to assumptions of unidimensionality and normality*, Appl. Psychol. Meas. 25 (2001), pp. 146–162.

[25] T. Micceri, *The unicorn, the normal curve, and other improbable creatures*, Psychol. Bull.105 (1989), pp. 156–166.

[26] D.C. Montgomery, *Design and Analysis of Experiments*, 6th ed., Chapman & Hall, London, 2004.

[27] R.J. Patz and B.W. Junker, *A straightforward approach to Markov chain Monte Carlo methods for item response models*, J. Educ. Behav. Stat. 24 (1999), pp. 146–178.

[28] R.J. Patz and B.W. Junker, *The applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses*, J. Educ. Behav. Stat. 24 (1999), pp. 342–366.

[29] A. Pewsey, *Problems of inference for Azzalini's skew-normal distribution*, J. Appl. Stat. 27 (2000), pp. 859–870.

[30] E.C. Poli, *Longitudinal study in Mathematics: Possibilities and reading of a reality of elementary school*, PhD diss. (in Portuguese), University of Campinas, 2007.

[31] S.K. Sahu, *Bayesian estimation and model choice in item response models*, J. Statist. Comput. Simul. 72 (2002), pp. 217–232.

[32] S.K. Sahu, D.K. Dey, and M.D. Branco, *A new class of multivariate skew distributions with applications to Bayesian regression models*, Canad. J. Stat. 31 (2003), pp. 129–150.

[33] F. Samejima, *Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling*, Psychometrika 62 (1997), pp. 471–493.

[34] J.R.S. Santos, *A multiple group item response theory model with skew normal latent trait distributions under centred parameterization*, Master diss. (in Portuguese), Department of Statistics, University of Campinas, 2012.

[35] J.R. Santos, C.L.N. Azevedo, and H. Bolfarine, *A multiple group item response theory model with centred skew normal latent trait distributions under a Bayesian framework, preprint* 2012. Available at http://www.ime.unicamp.br/sites/default/files/rel_pesq/rp10-12.pdf

[36] T.-J. Seong, *Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions*, Appl. Psychol. Meas. 14 (1990), pp. 299–311.

[37] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linden, *Bayesian measures of model complexity and fit*, J. R. Statist. Soc. 64(3) (2002), pp. 583–639.

[38] H. Swaminathan, R.K. Hambleton, S.G. Sireci, D. Xing, and S.M. Rizavi, *Small ample size estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates*, Appl. Psychol. Meas. 27 (2003), pp. 27–51.

[39] J.A. Wollack, D.M. Bolt, A.S. Cohen, and Y.-S. Lee, *Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo Estimation*, Appl. Psychol. Meas. 26 (2002), pp. 339–352.

Appendix

Using the augmented likelihood (6) and the prior distribution (19), we have that

$$
\begin{aligned}
p(\boldsymbol{\theta}_{..}, \boldsymbol{h}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_\theta, \boldsymbol{t}_{.} \mid \boldsymbol{y}_{...}, \boldsymbol{\eta}_{\boldsymbol{\zeta}}, \boldsymbol{\eta}_{\boldsymbol{\eta}}) &\propto p(\boldsymbol{Z}_{...} \mid \boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{y}_{...}) p(\boldsymbol{\theta}_{..} \mid \boldsymbol{h}_{..}, \boldsymbol{\eta}_\theta) p(\boldsymbol{\zeta} \mid \boldsymbol{\eta}_{\boldsymbol{\zeta}}) p(\boldsymbol{\eta}_\theta \mid \boldsymbol{t}_{.}, \boldsymbol{\eta}_{\boldsymbol{\eta}}) p(\boldsymbol{t}_{.}) \\
&= \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} \prod_{i \in I_{jk}} p(z_{ijk} \mid \theta_{jk}, \boldsymbol{\zeta}_i, y_{ijk}) \right\} \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} p(\theta_{jk} \mid h_{jk}, \boldsymbol{\eta}_{\theta_k}) \right\} \\
&\quad \times \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} p(h_{jk}) \right\} \left\{ \prod_{i=1}^{I} p(\boldsymbol{\zeta}_i) \right\} \left\{ \prod_{k=1}^{K} p(\eta_{\theta_k} \mid \boldsymbol{\eta}_{\eta_k}) \right\} \left\{ \prod_{k=1}^{K} p(t_k) \right\} \\
&\propto \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} \prod_{i \in I_{jk}} \exp\{-0.5(z_{ijk} - a_i \theta_{jk} + b_i)^2\} \mathbb{1}_{(Z_{ijk}, y_{ijk})} \right\} \\
&\quad \times \left\{ \prod_{k=1}^{K} \varsigma_{\theta_k}^{-\frac{n_k}{2}} \prod_{j=1}^{n_k} \exp\left\{ -\frac{1}{2\varsigma_{\theta_k}^2} \left[\theta_{jk} - \xi_{\theta_k} - \tau_{\theta_k} h_{jk} \right]^2 \right\} \right\} \\
&\quad \times \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} \exp\left(-\frac{1}{2} h_{jk}^2 \right) \mathbb{1}_{(h_{jk} > 0)} \right\} \\
&\quad \times \left\{ \prod_{i=1}^{I} \exp[-0.5(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta)^\top \boldsymbol{\Psi}_\zeta^{-1} (\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta)] \mathbb{1}_{(a_i > 0)} \right\} \\
&\quad \times \left\{ \prod_{k=1}^{K} \frac{1}{\varsigma_{\theta_k}^2} \exp\left(-\frac{1}{2} \frac{\tau_{\theta_k}^2 t_k}{\varsigma_{\theta_k}^2 \varphi^2} \right) t_k^{((d+1)/2)-1} \exp\left(-\frac{b}{2} t_k \right) \right\}, \qquad \text{(A1)}
\end{aligned}
$$

if we use the prior (16), or

$$
\begin{aligned}
p(\boldsymbol{\theta}_{..}, \boldsymbol{h}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_\theta \mid \boldsymbol{y}_{...}, \boldsymbol{\eta}_{\boldsymbol{\zeta}}, \boldsymbol{\eta}_{\boldsymbol{\eta}}) &\propto p(\boldsymbol{Z}_{...} \mid \boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{y}_{...}) p(\boldsymbol{\theta}_{..} \mid \boldsymbol{h}_{..}, \boldsymbol{\eta}_\theta) p(\boldsymbol{\zeta} \mid \boldsymbol{\eta}_{\boldsymbol{\zeta}}) p(\boldsymbol{\eta}_\theta \mid \boldsymbol{\eta}_{\boldsymbol{\eta}}) \\
&= \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} \prod_{i \in I_{jk}} p(z_{ijk} \mid \theta_{jk}, \boldsymbol{\zeta}_i, y_{ijk}) \right\} \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} p(\theta_{jk} \mid h_{jk}, \boldsymbol{\eta}_{\theta_k}) \right\} \\
&\quad \times \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} p(h_{jk}) \right\} \left\{ \prod_{i=1}^{I} p(\boldsymbol{\zeta}_i) \right\} \left\{ \prod_{k=1}^{K} p(\eta_{\theta_k} \mid \boldsymbol{\eta}_{\eta_k}) \right\} \left\{ \prod_{k=1}^{K} p(t_k) \right\} \\
&\propto \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} \prod_{i \in I_{jk}} \exp\left\{ -0.5(z_{ijk} - a_i \theta_{jk} + b_i)^2 \right\} \mathbb{1}_{(Z_{ijk}, y_{ijk})} \right\}
\end{aligned}
$$

$$\times \left\{ \prod_{k=1}^{K} \varsigma_{\theta_k}^{-n_k/2} \prod_{j=1}^{n_k} \exp\left\{ -\frac{1}{2\varsigma_{\theta_k}^2} [\theta_{jk} - \xi_{\theta_k} - \tau_{\theta_k} h_{jk}]^2 \right\} \right\}$$

$$\times \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_k} \exp\left(-\frac{1}{2} h_{jk}^2 \right) \mathbb{1}_{(h_{jk}>0)} \right\}$$

$$\times \left\{ \prod_{i=1}^{I} \exp\left[-0.5(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta)^\top \boldsymbol{\Psi}_\zeta{}^{-1} (\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta) \right] \mathbb{1}_{(a_i>0)} \right\}$$

$$\times \left\{ \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\beta}_\theta - \boldsymbol{\mu}_\theta \right)^\top \Sigma_\theta^{-1} \left(\boldsymbol{\beta}_\theta - \boldsymbol{\mu}_\theta \right) \right\} \right\}$$

$$\times \left\{ (\varsigma_{\theta_k}^2)^{-(\alpha_{\varsigma\theta}+1)} \exp\left(-\frac{\beta_{\varsigma\theta}}{\varsigma_{\theta_k}^2} \right) \right\}. \tag{A2}$$