Computational Statistics and Data Analysis 56 (2012) 4399–4412

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A Bayesian generalized multiple group IRT model with model-fit assessment tools

Caio L.N. Azevedo[a,*], Dalton F. Andrade[b], Jean-Paul Fox[c]

[a] Department of Statistics, University of Campinas, Brazil

[b] Department of Informatics and Statistics, Federal University of Santa Catarina, Brazil

[c] Department of Research Methodology, University of Twente, The Netherlands

ARTICLE INFO

Article history:
Received 2 June 2011
Received in revised form 23 March 2012
Accepted 24 March 2012
Available online 1 April 2012

Keywords:
Multiple groups
Gibbs sampling
Posterior
Predictive checking
Bayesian residual analysis

ABSTRACT

The multiple group IRT model (MGM) proposed by Bock and Zimowski (1997) provides a useful framework for analyzing item response data from clustered respondents. In the MGM, the selected groups of respondents are of specific interest such that group-specific population distributions need to be defined. The main goal is to explore the potentials of an MCMC estimation procedure and Bayesian model-fit tools for the MGM. We develop a full Gibbs sampling algorithm (FGSA) for estimation as well as a Metropolis-Hastings within Gibbs sampling algorithm (MHWGS) in order to use non-conjugate priors. The FGSA is compared with Bilog–MG, which uses marginal maximum likelihood (MML) and marginal maximum a posteriori (MMAP) methods. That is; Bilog–MG provides maximum likelihood (ML) and expected a posteriori (EAP) estimates for both item and population parameters, and maximum a posteriori (MAP) estimates for the latent traits. We conclude that, in general, the results from our approach are slightly better than Bilog–MG. Besides a simultaneous MCMC estimation procedure, model-fit assessment tools are developed. Furthermore, the prior sensitivity is investigated with respect to the parameters of the latent population distributions. It will be shown that the FGSA provides a wide set of model-fit tools. The proposed model assessment tools can evaluate important model assumptions of (1) the item response function (IRF) and (2) the latent trait distributions. The utility of the proposed estimation and model-fit assessment methods will be shown using data from a longitudinal data study concerning first to fourth graders of sampled Brazilian public schools.

1. Introduction

In educational assessment, clinical trials and bio essays among other fields, it is common to observe examinees (subjects) from different groups. The groups can be characterized by gender, grade, social level, and so on. The group heterogeneity can reflect different behaviors. Therefore, it is important to take such heterogeneity into account. Attention will be focused on applications where the number of groups is limited and/or there is a specific interest in the sampled groups. The population distribution representing the clustered respondents completely specifies the distribution of respondents in each group, and no assumptions will be made about groups that are not selected. Then, inferences can be made with respect to the sampled groups but not to some higher level of population of groups.

Bock and Zimowski (1997) developed an IRT model where each group has a specific latent trait distribution. This multiple group model (MGM) has an additional set of parameters: multiple population parameters, which characterize the latent

* Corresponding author. Tel.: +55 19 35216060.
E-mail address: cnaber@ime.unicamp.br (C.L.N. Azevedo).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.03.017

population distributions. Bock and Zimowski (1997) developed a frequentist-based approach for parameter estimation and model assessment. Furthermore, they suggested a nonparametric estimation method of Mislevy (1984) for the latent trait distributions. The approach performs the equating process simultaneously with the estimation of item and population parameters. Subsequently, the latent traits can be estimated using maximum likelihood or Bayesian methods. In this way, the results tend to be more accurate than that ones obtained by other equating techniques, see Andrade et al. (2000) and Kolen and Brennan (2004). The MGM has been studied in different applications. Béguin and Glas (2001) developed a multidimensional IRT model and a Gibbs sampling algorithm for a multiple groups three parameter model and studied some methods of model fit. Gonçalves (2006) proposed a new multiple groups model for DIF assessment and an estimation method based on a Metropolis–Hastings within Gibbs sampling algorithm. Other discussions about MGM can be found in Bock and Zimowski (1997).

Other works had considered an IRT analysis for multiple groups framework without using the MGM. For example, Kim et al. (2005) and Múthen and Lehman (1985) studied DIF detection and its effects on the parameter estimation. Glas (1998) and Penfield (2001) developed statistical tests for DIF assessment within the same context. Fox and Glas (2001) proposed a multilevel model to take into account the similarities among subjects that belong to the same group. In their approach, the population distribution represents a population of groups, and inferences are made with respect to the population of groups. This approach is different from the framework considered by Bock and Zimowski (1997), where interest is focused on the sampled groups.

The MGM will be generalized to handle response data from groups with different latent trait variances, incomplete designs where the number of respondents and/or the number of test items varies across groups and/or respondents, and more flexible population distributions when the normal distribution assumptions do not hold. This includes heavier tails distributions, as the Student's t, and asymmetric ones as the skew normal. This generalized MGM framework also contains a wide set of response functions, which includes the probit and logit functions. A fully Bayesian analysis framework using Gibbs sampling (see Gamerman and Lopes, 2006) and Metropolis-Hasting within-Gibbs sampling algorithms will be proposed to handle the aforementioned aspects and shortcomings of the present statistical methods.

After introducing the Bayesian generalized MGM, the accuracy of the MCMC estimation methods as well as the prior sensitivity are assessed using simulation studies. Results are compared with Bilog–MG, which produces expected a posteriori estimates. Then, the generalized MGM is illustrated using a real data study. The data set consists of a longitudinal study of children from the first grade up to fourth grade of various Brazilian public schools. The student achievements, at different grade levels, are estimated simultaneously and provide information about student level of achievement and grade level of achievement. The proposed model assessment tools are used to evaluate the fit of the MGM to the real data. In the last section, the results and some possible model extensions are discussed.

2. The multiple group model

One or more different tests are administered by the (randomly selected) examinees of each group. The tests have common items and the structure can be recognized as an incomplete block design, see Montgomery (2004). We will assume that each group has a reasonable number of subjects. In summary, we are dealing with a set of n examinees clustered in K groups, with n_k examinees in group k, and $n = \sum_{k=1}^{K} n_k$. The examinees of each 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 examinee j ($j = 1, \ldots, n_k,$) belonging to group k ($k = 1, \ldots, K$), $\boldsymbol{\theta}_{\cdot k} = (\theta_{j1}, \ldots, \theta_{jK})^t$ is the vector of the latent traits of the examinees of the group k, and $\boldsymbol{\theta}_{\cdot\cdot} = (\boldsymbol{\theta}_{.1}, \ldots, \boldsymbol{\theta}_{\cdot K})^t$ is the vector with all latent traits; Y_{ijk} is the response of the examinee j, of the group k to the item i ($i = 1, \ldots, I$), $\boldsymbol{Y}_{\cdot jk} = (Y_{1jk}, \ldots, Y_{Ijk})^t$ is the response vector of the examinee j of the group k, $\boldsymbol{Y}_{\cdot\cdot k} = (\boldsymbol{Y}_{\cdot 1k}{}^t, \ldots, \boldsymbol{Y}_{\cdot n_k k}{}^t)^t$ is the response vector of all examinees of the group k, $\boldsymbol{Y}_{\ldots} = (\boldsymbol{Y}_{\cdot 1\cdot}{}^t, \ldots, \boldsymbol{Y}_{\cdot n_k\cdot}{}^t)^t$ is the whole response set and $(\boldsymbol{y}_{\cdot 1\cdot}{}^t, \ldots, \boldsymbol{y}_{\cdot n_k\cdot}{}^t)^t$ are the observed values; $\boldsymbol{\zeta}_i$ is the vector of parameters of the item i, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1{}^t, \ldots, \boldsymbol{\zeta}_I{}^t)^t$ is the whole set of item parameters, $\boldsymbol{\eta}_{\theta_k}$ is the vector with the population parameters k and $\boldsymbol{\eta}_{\boldsymbol{\theta}} = (\boldsymbol{\eta}_{\theta_1}{}^t, \ldots, \boldsymbol{\eta}_{\theta_K}{}^t)^t$ is the whole set of population parameters.

When modeling the grouping structure of subjects using group-specific normal population distributions for the latent traits, the MGM can be seen as a natural extension of the two-parameter (probit) item response model, which is represented by,

$$Y_{ijk} \mid (\theta_{jk}, \boldsymbol{\zeta}_i) \sim \text{Bernoulli}(P_{ijk})$$

$$P_{ijk} = P(Y_{ijk} = 1 \mid \theta_{jk}, \boldsymbol{\zeta}_i) = \Phi(a_i\theta_{jk} - b_i) \tag{1}$$

$$\theta_{jk} \mid \boldsymbol{\eta}_{\theta_k} \sim N(\mu_{\theta_k}, \psi_{\theta_k}), \tag{2}$$

where $\Phi(\cdot)$ stands for the cumulative normal function. In this parametrization, the difficulty parameter $b_i = a_i b_i^*$ is a transformation of the commonly used difficulty parameter denoted as b_i^*.

In the literature, other item response functions (IRFs), such as the skew probit, logit, or log–log are considered as well as other latent trait population distributions. To include this flexibility in response functions and latent trait distributions, a generalized MGM is defined as a mixture (indexed $l = 1, \ldots, L$) of different response functions based on different cumulative distribution functions (indexed $h = 1, \ldots, H$), and different latent trait distributions across items and groups of subjects. Then, the success probability is stated as

$$P_{ijk} = P(Y_{ijk} = 1 \mid \theta_{jk}, \boldsymbol{\zeta}_i, \boldsymbol{\phi}) = \sum_{l=1}^{L} \prod_{h=1}^{H} F_{lh} \left(\theta_{jk}, \boldsymbol{\zeta}_i, \boldsymbol{\phi}\right) \tag{3}$$

$$\theta_{jk} \mid \boldsymbol{\eta}_{\theta_k} \sim D(\boldsymbol{\eta}_\theta), \tag{4}$$

where $F(\cdot)$ represents a cumulative distribution function with parameters, $\boldsymbol{\phi} \cdot D(\cdot)$ represents a continuous population distribution function, where $\boldsymbol{\eta}_\theta$ denote the population parameters.

This modeling framework comprehends the well-known one-, two- and three-parameter item response models using a probit, logit, log–log, Student's t, skew probit, or skew probit link function, among others. In addition, the latent trait population distribution can be characterized by a normal, Student's t, skew normal, skew Student's t or finite mixture of normals, among others.

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. In the same way, the MGM for mixed response data will contain response models for discrete binary and polytomous response data.

In terms of latent response augmented data, following Albert (1992) and/or Fox (2010), the one- and two-parameter MGM with explanatory variables at different hierarchical levels is given by,

$$Z_{ijk} = \boldsymbol{X}_{ijk}\boldsymbol{\beta}_{ik} + a_{ik}\theta_{jk} - b_{ik} + \xi_{ijk} \tag{5}$$

$$\theta_{jk} = \boldsymbol{W}_{jk}\boldsymbol{\gamma}_k + \epsilon_{jk} \tag{6}$$

$$\boldsymbol{\gamma}_k = \boldsymbol{T}_k\boldsymbol{\tau} + \boldsymbol{u}_k, \tag{7}$$

where the prior parameters are conditionally modeled as

$$\ln a_{ik} = \boldsymbol{H}_{ik}^{(a)}\boldsymbol{\delta}_k^{(a)} + \epsilon_{ik}^{(a)} \tag{8}$$

$$b_{ik} = \boldsymbol{H}_{ik}^{(b)}\boldsymbol{\delta}_k^{(b)} + \epsilon_{ik}^{(b)}. \tag{9}$$

This conditional framework includes the modeling of DIF (differential item functioning), where the DIF is modeled in terms of non-invariant item response functions and/or non-invariant discrimination and difficulty parameters. Furthermore, collateral information can be used to model the item response functions and/or the latent trait population distribution. Furthermore, heterogeneity in large-scale comparative research, and latent trait multilevel distributions can be conditionally modeled using explanatory information at the item, subject, and group level. The models proposed by Azevedo et al. (2011), da-Silva and Gomes (2011), Fox and Glas (2001), and Soares et al. (2009) can be viewed as particular cases of this general conditional MGM framework. Some comments concerning the simultaneous estimation of the generalized MGM model will be discussed in the Appendix. Also, comments on identification issues are provided in the Appendix. Below, attention is focused on the MGM represented in Eqs. (1) and (2).

2.1. Priors and hyperpriors

Hierarchical priors are specified to control the prior parameters and to let the amount of shrinkage be estimated from the data. The group-specific population distribution consists of a mean and variance parameter. For the reference group, the mean and variance parameters are fixed to zero and one, respectively. For the other groups, an exchangeable normal inverse-gamma distribution is specified for the mean and variance parameter. The mean and variance parameters of the non-reference groups are assumed to be conditionally independent distributed. The values of the hyperparameters of the normal inverse-gamma distribution define the shrinkage towards a general population mean across the fixed groups.

For the item parameters an exchangeable truncated multivariate normal distribution is considered.

3. MCMC estimation for the MGM

An augmented data scheme is introduced to sample continuous normally distributed item response data, denoted as **Z**, given discrete observed item response data, denoted as **y**. Following Albert (1992),

$$Z_{ijk} \mid (\theta_{jk}, \boldsymbol{\zeta}_i, Y_{ijk}) \sim N(a_i\theta_{jk} - b_i, 1), \tag{10}$$

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

To handle an incomplete block design, an indicator variable **I** is defined that defines the set of administered items according to the design. For each administered item response information is recorded. This indicator variable is described by,

$$I_{ijk} = \begin{cases} 1, & \text{item } i \text{ administered for examinee } j \text{ of group } k \\ 0, & \text{missing by design.} \end{cases} \tag{11}$$

The indicator matrix **I** describes the patterns of missing data that is deliberately allowed to be missing. These missing data are missing by design.

In the same way 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 is defined as,

$$V_{ijk} = \begin{cases} 1, & \text{observed response of examinee } j \text{ of group } k \text{ on item } i \\ 0, & \text{missing response.} \end{cases} \tag{12}$$

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

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

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

$$\begin{aligned} & p(\boldsymbol{\theta}_{\ldots}, \boldsymbol{\zeta}, \boldsymbol{\eta}_\theta \mid \boldsymbol{y}_{\ldots}, \boldsymbol{v}_{\ldots}, \boldsymbol{z}_{\ldots}) \\ & \quad = p(\boldsymbol{\theta}_{\ldots}, \boldsymbol{\zeta}, \boldsymbol{\eta}_\theta \mid \boldsymbol{y}_{\ldots}, \boldsymbol{z}_{\ldots}) \\ & \quad \propto \prod_{k=1}^{K} \left[\prod_{j=1}^{n_k} \left[\left[\prod_{i \in I_{jk}} p\left(z_{ijk} \mid \theta_{jk}, \boldsymbol{\zeta}_i, y_{ijk}\right) \right] p(\theta_{jk} \mid \boldsymbol{\eta}_{\theta_k}) \right] \times p(\boldsymbol{\eta}_{\theta_k} \mid \boldsymbol{\eta}_\eta) \right] \prod_{i=1}^{I} p(\boldsymbol{\zeta}_i \mid \boldsymbol{\eta}_\zeta), \end{aligned} \tag{13}$$

where $\boldsymbol{\eta}_\zeta = (\boldsymbol{\mu}_\zeta, \boldsymbol{\Psi}_\zeta)$ and $\boldsymbol{\eta}_\eta = (\mu_0, \nu_0, \kappa_0)$ are the hyperparameters associated with $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}_{\boldsymbol{\theta}}$, respectively.

In the Appendix it is shown that the full conditionals are known and easy to sample from. A Gibbs sampling scheme is defined that describes a sampling procedure consisting of six steps. Let $(\cdot)$ denote the set of all necessary parameters, the steps of the Gibbs sampling scheme is summarized as follows;

1. Start the algorithm by choosing suitable initial values.
   Repeat steps 2–6:.
2. Simulate Z_{ijk} from $Z_{ijk} \mid (\cdot), k = 1, \ldots, K, i = 1, \ldots, I, j = 1, \ldots, n_k$.
3. Simulate θ_{jk} from $\theta_{jk} \mid (\cdot), k = 1, \ldots, K, j = 1, \ldots, n_k$.
4. Simulate $\boldsymbol{\zeta}_i$ from $\boldsymbol{\zeta}_i \mid (\cdot), i = 1, \ldots, I$.
5. Simulate μ_{θ_k} from $\mu_{\theta_k} \mid (\cdot), k = 1, \ldots, K$.
6. Simulate ψ_{θ_k} from $\psi_{\theta_k} \mid (\cdot), k = 1, \ldots, K$.

The above mentioned full Gibbs sampling scheme will be referred to as MCMC scheme 1. This MCMC scheme will be modified in three different ways.

The six-step Gibbs sampler can be modified by blocking several highly correlated parameters. The blocking of highly correlated parameters can improve the mixing and convergence properties of the MCMC chains. The parameters μ_{θ_k} and ψ_{θ_k} can easily be sampled simultaneously from a normal inverse-gamma density. Note that the population parameters are assumed to be conditionally independent across groups, which makes it uninteresting to sample jointly all mean and variance population parameters. Other blocked sampling steps are possible but not in a full Gibbs sampling approach. Furthermore, Metropolis–Hastings steps are required that can evaluate proposals of other blocked parameters. Such an implementation depends on the quality of proposal distributions for the blocked parameters, which are not known for the MGM parameters and beyond the scope of the present paper.

MCMC scheme 2 is defined to handle non-conjugate priors using a Metropolis–Hastings within Gibbs algorithm. The normal and a lognormal prior for the difficulty and discrimination parameters, respectively, are popular priors. The priors are used in Bilog-MG, where maximum a posteriori (MAP) item parameter estimates are computed.

Note that the full Gibbs sampling (MCMC) scheme 1 is designed for the MGM model with a conjugate item prior. The Metropolis–Hastings within Gibbs sampling algorithms, MCMC scheme 2, is designed for the MGM with a non-conjugate item prior.

In addition to the Full Gibbs sampling and Metropolis–Hastings within Gibbs sampling algorithms, in the Appendix we discuss alternative algorithms. One version, proposed by Gonzalez (2004), is based on the joint simulation of augmented data and latent traits, see Chen et al. (2000).

The MGM can be identified by fixing the mean and the variance of the latent traits distribution of the reference group as in Soares et al. (2009), since the tests are linked by anchor items.

4. Simulation study

In this section, MGMs will be compared that differ with respect to prior specifications. MGMs with conjugate and non-conjugate item priors and different degrees of hyperprior information are considered. Furthermore, a comparison will be made with an MGM specified in Bilog-MG (Zimowski et al., 1996).

A total of ten replicated data sets were generated, which was based on the work of Azevedo and Andrade (2010) and De Ayala and Sava-Bolesta (1999). The simulated groups consisted of 500 subjects. A linked design was defined for 60 items such that 120 item parameters had to be estimated and 1,500 person parameters. The general population model for the person parameters led to an additional set of 6 parameters. In this simulation study, 1,000 responses per item were simulated.

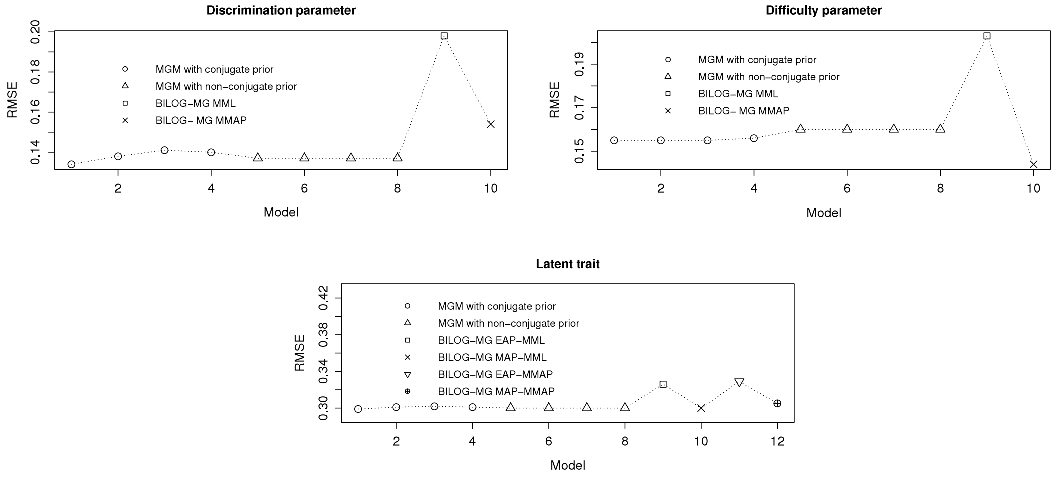

Fig. 1. RMSE estimates for discrimination, difficulty, and latent trait parameters across ten replicated data sets.

The latent traits were sampled from group-specific independent normal distributions, with group means $\boldsymbol{\mu}_\theta = (0, 1, 2)$ and group variances $\boldsymbol{\psi}_\theta = (1, 1.1, 0.9)$. Note that the characteristics of the real data resembles those of the simulated data.

Four MGMs were specified with different hyperparameter values. For the item parameters, the hyperparameters were set at $\mu_\zeta = (1, 0)$ and $\boldsymbol{\Psi}_\zeta = \mathrm{diag}(0.5, 3)$. For the latent trait parameters, the hyperparameters (ν_0, κ_0) were varied. For model MGM 1, a noninformative (improper) population prior $(0, 0)$ was defined. For MGM 2, a weak noninformative (proper) population prior $(.001, .001)$ was defined. For MGM 3 and MGM 4 a moderate informative population prior $(4.02, 2.02)$ and a strong informative population prior $(6, 4)$ was defined, respectively. The general population mean was always fixed at zero, $(\mu_0 = 0)$. Each MGM was specified with a conjugate and non-conjugate item prior. As a result, for each replicated data set eight different MGMs were estimated.

Bilog–MG was used to compute marginal maximum likelihood (MML) and marginal maximum a posteriori estimates (MMAP) (see, Bock and Aitkin, 1981; Mislevy, 1986). The default priors were used. Bilog–MG does not allow to specify hyperpriors for the item population parameters. Two different latent trait estimates were considered from Bilog–MG, the expected a posteriori (EAP) and the maximum a posteriori (MAP) estimate. For each set of item parameter estimates (MML or MMAP), the EAP and MAP latent trait estimates were computed. For the latent trait estimates obtained through BILOG-MG, the following nomenclature was used in the Figures: EAP–MML, MAP–MML, EAP–MMAP and MAP–MMAP, respectively.

In order to asses the parameter recovery, two different statistics were defined. Let ϑ be the parameter of interest. Let $\widehat{\vartheta}_r$ be the posterior mean estimate related to data set r, and $\widehat{\widehat{\vartheta}}$ the general posterior mean estimate, which is the average of the R posterior means, each related to one of the R data sets. Now, a root mean square error (RMSE), and the absolute value of the relative bias (AVRB) are defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\vartheta - \widehat{\vartheta}_r\right)^2},$$

and

$$\mathrm{AVRB} = \frac{|\widehat{\widehat{\vartheta}} - \vartheta|}{|\vartheta|},$$

respectively.

In Fig. 1, the estimated RMSEs are plotted based on ten replicated data sets. A RMSE value was computed for the discrimination parameters, for the difficulty parameters, and for the latent trait parameters. In each subplot, the four circles correspond to the RMSE estimates of the MGM with noninformative, weak, moderate, and strong prior information using the conjugate item prior. The four triangles correspond to the RMSE estimates of the four MGMs with the non-conjugate item prior. The RMSE estimates using MML and MMAP item parameter estimates from Bilog–MG are plotted. Furthermore, the RMSE estimates are plotted based on EAP and MAP latent trait estimates using MML and MMAP item parameter estimates.

In general, the RMSEs of the MGMs with non-conjugate item priors show more accurate results for the discrimination and latent trait parameters than the other MGMs. The MGMs with conjugate item priors show slightly better results for the difficulty parameters. The different hyperprior settings hardly influenced the RMSE. It was expected that the posterior population means will shrink more towards zero when using the informative hyperpriors in comparison to the noninformative priors, since the hyperprior specification will induce bias for the groups with a non-zero population mean. In the present setting, the hyperprior information differences did not lead to substantial differences in posterior mean

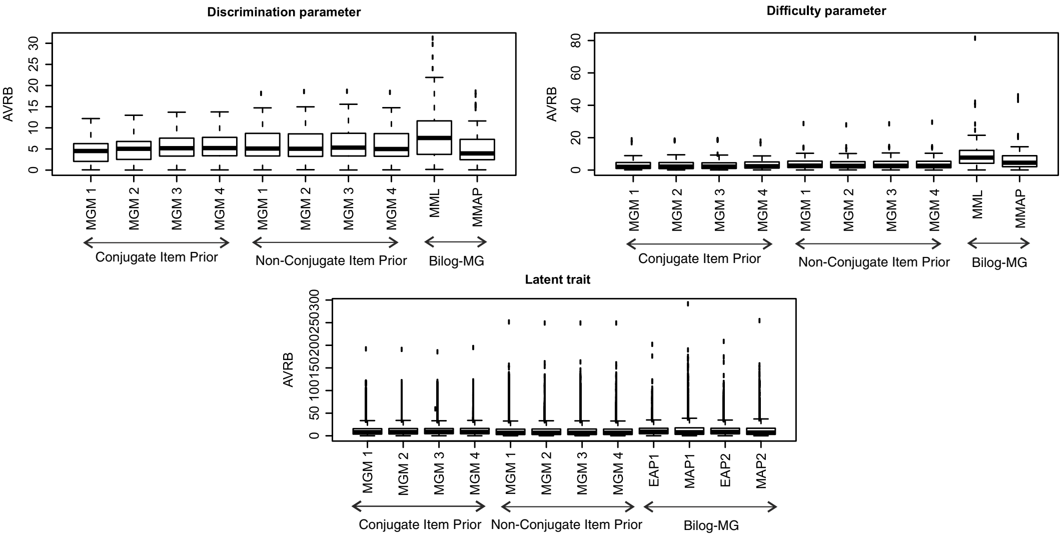

Fig. 2. AVRB estimates for the discrimination, difficulty, and latent trait parameters across ten replicated data sets.

estimates. For the MGMs with conjugate and non-conjugate item prior, only a relatively small increase in RMSE was detected when increasing the hyperprior information. It can be concluded that the hyperprior specification supports the multiple group structure in such a way that the data can be used to estimate the prior population parameters.

Fig. 2 shows comparable results, the absolute value of the relative bias is smaller for the MGMs with hyperprior specifications compared to the MGM specified in Bilog–MG. However, the MGMs with a conjugate item prior method shows slightly better results than the MGMs with the non-conjugate item prior. The amount of bias does not vary over the different hyperprior settings. Only a slight increase in bias can be detected when using the more informative hyperpriors. It can be concluded that a misspecification of the hyperprior parameters will not lead to substantial bias in the posterior mean estimates. The simulation study shows that the MGMs with conjugate and non-conjugate item priors perform well and show more accurate results than Bilog–MG model with respect to the used statistics.

5. MGM model assessment: Posterior predictive checks

The fit of the MGM can be evaluated using different Bayesian posterior predictive tests and Bayesian residual analysis techniques. The literature about posterior predictive checks for Bayesian item response models shows several diagnostics for evaluating the model fit. A general discussion about posterior predictive checks for IRT models can be found in, among others, (Stern and Sinharay, 2005; Sinharay, 2006; Sinharay et al., 2006; Fox, 2004; Fox and Glas, 2005; Fox, 2010). In this section, it is shown that the posterior predictive tests can be generalized to make them applicable for the MGM. Each posterior predictive test is based on a discrepancy measure. The discrepancy measure is defined in such a way that a specific assumption or the general fit of the model can be evaluated. The main idea is to generalize the well known discrepancy measures to a multiple group structure. Furthermore, most discrepancy measures are based on residuals, where usually Pearson residuals are used. A generalization will be made such that discrepancy measures can be based on other types of residuals, which will lead to a wide variety of model diagnostics.

Let $\boldsymbol{y}^{obs}$ be the matrix of observed responses, and $\boldsymbol{y}^{rep}$ the matrix of replicated responses generated from its posterior predictive distribution. The posterior predictive distribution of the response data of group k is represented by

$$p\left(\boldsymbol{y}_k^{rep} \mid \boldsymbol{y}_k^{obs}\right) = \int p\left(\boldsymbol{y}_k^{rep} \mid \boldsymbol{\vartheta}_k\right) p\left(\boldsymbol{\vartheta}_k \mid \boldsymbol{y}_k^{obs}\right) d\boldsymbol{\vartheta}_k,$$

where $\boldsymbol{\vartheta}_k$ denotes the set of model parameters corresponding to group k. Generally, given a discrepancy measure $D\left(\boldsymbol{y}_k, \boldsymbol{\vartheta}_k\right)$, the replicated data can be used to evaluate whether the discrepancy value given the observed data is typical under the model. A p-value can be defined that quantifies the extremeness of the observed discrepancy value in group k,

$$p_0\left(\boldsymbol{y}_k^{(obs)}\right) = P\left(D\left(\boldsymbol{y}_k^{(obs)}, \boldsymbol{\vartheta}_k\right) \geq D\left(\boldsymbol{y}_k^{(rep)}, \boldsymbol{\vartheta}_k\right) \middle| \boldsymbol{y}_k^{(obs)}\right),$$

where the probability is taken over the joint posterior of $(\boldsymbol{y}_k^{(rep)}, \boldsymbol{\vartheta}_k)$. In some cases, the discrepancy measure can be generalized from the group level to the population level. In that case, the discrepancy measure can be used to evaluate model fit at the group and population level.

5.1. Residual-based discrepancy measures

The discrepancy measures for posterior predictive checks are often based on Pearson residuals. This type of residual represents the difference between the observed and expected value scaled by the standard deviation of the observation. A discrepancy measure $D_p\,(y)$ can be defined as the squared residual, which corresponds to the well-known Pearson goodness-of-fit statistic. Such discrepancy measures can be generalized by considering other types of residuals. For generalized linear models, different types of residuals were discussed by McCullagh and Nelder (1989).

The most straightforward residual is the response residual, which is simply the difference between the observation and its expected value. However, this residual is usually not used in model diagnostics. The deviance residual has the nice feature that it connects to the likelihood contribution of the corresponding observation. The discrepancy measure based on deviance residual is denoted by $D_v\,(y)$ and represents the likelihood contribution of the observation with a positive (negative) sign when the observation is greater (smaller) than its expected value. The squared sum of deviance residuals is equal to the deviance of the model.

Anscombe residuals were proposed such that the resulting residuals had a distribution that was approximately normal. In the present context, this specific feature of approximately normally distributed residuals is not essential, since discrepancy measures can be evaluated using posterior predictive data. Besides, the evaluation of the posterior predictive check will be much more complicated when using discrepancy measures based on Anscombe residuals. This also holds for the likelihood residual, which is a combination of the standardized Pearson and standardized deviance residual.

Béguin and Glas (2001) proposed a posterior predictive check to evaluate the observed score distribution with the posterior predictive score distribution. For the MGM, the observed score distribution can be evaluated per group. For the Pearson residual, the generalized group-specific discrepancy measure is defined as,

$$D_p\,(\boldsymbol{y}_k) = \sum_l \frac{\left(n_{l,k} - E(n_{l,k})\right)^2}{V\,(n_{l,k})}, \tag{14}$$

where $n_{l,k}$ is the number of persons with l correct scores in group k. The $E(\cdot)$ and $V(\cdot)$ stand for the expectation and the variance, respectively, and computational details are given in the Appendix.

For the response residual, this discrepancy measure can be defined as

$$D_r\,(\boldsymbol{y}_k) = \sum_l \left(n_{l,k} - E(n_{l,k})\right)^2,$$

and for deviance residual,

$$D_v\,(\boldsymbol{y}_k) = \sum_l d(n_{l,k}),$$

where $d(n_{l,k})$ represents the likelihood contribution of the observation. The deviance residual is based on the deviance function represented by a constant minus twice the log-likelihood. The posterior predictive test will assess the extremeness of the observed deviance residuals under the model. Without the necessity to compare the model with a baseline model, the constant is set to zero and the term minus two cancels out in the posterior predictive evaluation. In a more general approach, it is possible to consider deviance residuals with respect to a baseline model by defining the constant as the log-likelihood of the baseline model.

Let $\Omega\,(n_{l,k})$ represent the set of subjects in group k with a score l, and let l_j denote the set of l items that were correctly scored by subject j. Then, the (conditional) likelihood contribution $d(n_{l,k})$ is represented by

$$p\left(\boldsymbol{y}_k \mid n_{l,k}, \boldsymbol{\theta}_k, \boldsymbol{\zeta}\right) = \prod_{j\in\Omega(n_{l,k})} \left[\prod_{i\in l_j} P_{ijk}^{Y_{ijk}} \prod_{i\notin l_j} \left(1 - P_{ijk}\right)^{(1-Y_{ijk})}\right].$$

Note that the posterior predictive p-value given the deviance discrepancy measure requires computing the posterior probability over the joint posterior of the replicated data and the person and item parameters. The discrepancy measure based on the deviance residual evaluates the fit of the model to the data in group k taking the likelihood contribution into account. In practice, the log-likelihood contribution of the observation can also be used, such that the deviance is represented by the log-likelihood.

Further on, discrepancy measures are proposed that are based on Pearson residuals, but they can also be defined for the response and the deviance residual.

The group-specific discrepancy measure for evaluating the observed score distribution can be defined at the population level,

$$D\,(\boldsymbol{y}) = \sum_k D\,(\boldsymbol{y}_k) = \sum_k \sum_l \frac{\left(n_{l,k} - E(n_{l,k})\right)^2}{V\,(n_{l,k})}, \tag{15}$$

by summing over the group-specific discrepancy measures. The group-specific measure provides information about model fit at the group level and the sum of group-specific measures provides information of the model fit over groups. Note that this discrepancy measure can also be used for polytomous response data.

In a similar way, the observed item-score distribution can be evaluated per group. Therefore, a group-specific item-based discrepancy measure can be defined as

$$D_{i,k}\left(\boldsymbol{y}_{ik}\right) = \sum_{l} \frac{\left(n_{l,ik} - E\left(n_{l,ik}\right)\right)^2}{V(n_{l,ik})}, \tag{16}$$

where $n_{l,ik}$ is the number of respondents in group k with l correct scores, and item i scored correctly. The discrepancy measure can be generalized by summarizing over groups,

$$D_i\left(\boldsymbol{y}_i\right) = \sum_{k}\sum_{l} \frac{\left(n_{l,ik} - E\left(n_{l,ik}\right)\right)^2}{V(n_{l,ik})} \tag{17}$$

and over items

$$D\left(\boldsymbol{y}\right) = \sum_{i}\sum_{k}\sum_{l} \frac{\left(n_{l,ik} - E\left(n_{l,ik}\right)\right)^2}{V(n_{l,ik})}. \tag{18}$$

Bayesian IRT residual analysis tools can be used to evaluate the fit of the MGM. In Fox (2010), an overview is given of Bayesian residuals and Bayesian latent residuals. The Bayesian residuals are easily computed as the difference between the observations and the expected values under the model. The marginal posterior distribution of each Bayesian residual can be computed and used to evaluate its extremeness. Note that the posterior variances of the Bayesian residuals differ and the residuals' posterior densities are therefore not directly comparable.

The residual analysis for the MGM can be extended by focusing on the latent within-group residual variation. The latent traits are assumed to be normally distributed within each group. Therefore, a latent residual can be defined as

$$r_{\theta_{jk}} = \frac{\theta_{jk} - \mu_{\theta_k}}{\sqrt{\psi_{\theta_k}}}, \tag{19}$$

where $r_{\theta_{jk}}$ is standard normally distributed given the latent mean and variance of group k. Now, a marginal posterior outlying probability that the absolute latent residual is larger than a constant c can be defined as

$$\begin{aligned}
P\left(|r_{\theta_{jk}}| > c \mid \boldsymbol{y}_k\right) &= \int\int\int P\left(|r_{\theta_{jk}}| > c \mid \theta_{jk}, \mu_{\theta_k}, \psi_{\theta_k}, \boldsymbol{y}_k\right) \times p\left(\theta_{jk}, \mu_{\theta_k}, \psi_{\theta_k} \mid \boldsymbol{y}_k\right) d\theta_{jk}\, d\mu_{\theta_k}\, d\psi_{\theta_k} \\
&= \int\int \Phi\left(|r_{\theta_{jk}}| > c \mid \theta_{jk}, \mu_{\theta_k}, \psi_{\theta_k}, \boldsymbol{y}_k\right) \times p\left(\theta_{jk}, \mu_{\theta_k}, \psi_{\theta_k} \mid \boldsymbol{y}_k\right) d\theta_{jk}\, d\mu_{\theta_k}\, d\psi_{\theta_k} \\
&\approx M^{-1} \sum_{m} \left[\Phi\left(r_{\theta_{jk}}^{(m)} > c \mid \theta_{jk}^{(m)}, \mu_{\theta_k}^{(m)}, \psi_{\theta_k}^{(m)}, \boldsymbol{y}_k\right) + \Phi\left(r_{\theta_{jk}}^{(m)} < -c \mid \theta_{jk}^{(m)}, \mu_{\theta_k}^{(m)}, \psi_{\theta_k}^{(m)}, \boldsymbol{y}_k\right)\right],
\end{aligned}$$

using M MCMC samples from the joint posterior. The estimated outlying probabilities are evaluated to identify extreme residual values. A latent trait is considered an outlier when its corresponding residual outlying probability is large. Notice that under the assumption of Gaussian distribution for the latent traits the Anscombe, deviance, and Pearson residuals are equivalent.

In each iteration, the assumption of normality of the latent residuals can be tested using a Kolmogorov–Smirnov test and evaluating normal quantile–quantile plots. The corresponding p-values can be calculated in each iteration and averaged over the iterations. The averaged p-values provide information about the normality assumption of the latent residuals.

6. A grade comparison of Brazilian public schools

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), regarding Mathematical achievement. Several countries participate in this study, including Brazil, 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 organization of curriculum. 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 in 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

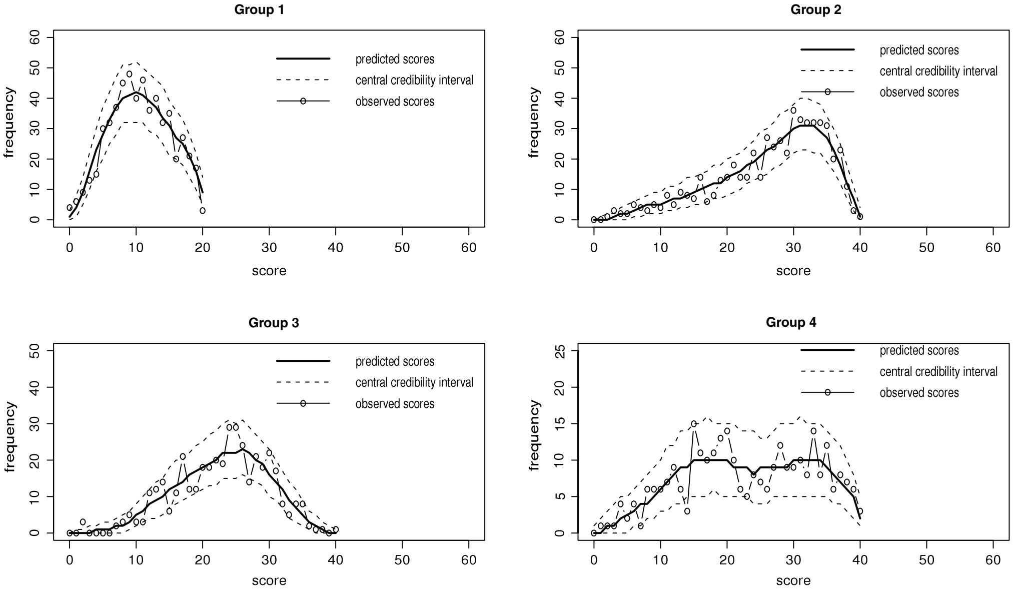

Fig. 3. Predictive and observed score distributions for each grade year.

in schools. Several student-level and teacher-level background variables (such as gender, skin color, age, education level of the parents) were collected.

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 responded to a test of 20 items, the second graders to 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 to 80 items, including the third-grade items, respectively. All items are were corrected as right-wrong items.

For grade two till four, 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 one 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. The data set is available upon request from the authors.

6.1. Parameter estimates and model fit

The MGM, where each group represents a grade group, was fitted, where MCMC scheme 1, see Section 3, was used to estimate all parameters. In several steps, the fit of the MGM was evaluated.

The overall fit of the model was evaluated using the general discrepancy measure, which led to a p-value of.103. This indicates that it cannot be concluded that the MGM does not fit the data. In Fig. 3, the observed score distribution (labeled observed scores) is compared with the estimated score distribution (labeled predicted scores) in each grade year. It can be seen in each grade that the observed scores are close to the predicted scores and are located within the credible region. Note that the predicted score distributions are posterior predictive distributions and they depend on the posterior distributions of the model parameters. Quantile–quantile plots of the Pearson residuals (not shown), calculated using the EAP estimates of the parameters, suggest that the normality assumption is reasonable for group 3. There is an indication of a slight departure from the normality assumption (heavy tails and asymmetry). This indicates that a model with non Gaussian latent traits distribution should be considered, see Azevedo et al. (2011). However, the observed values of the testing statistic of the Kolmogorov–Smirnov test indicate that the normality assumption should not be rejected in any group.

The results of the item-fit analysis are shown in Fig. 4. The item-based discrepancy measure, based on Pearson (denoted as chi-square statistic) and deviance residuals (denoted as deviance statistic), was used to evaluate the fit of the items. When looking at the chi-square statistic, it can be seen that most items fit well except for some items that were administered to grade-one students. For each item, the log-likelihoods of the response observations are not that extreme, which leads to moderate p-values for the deviance statistic.

The items with low p-values do not discriminate very well, which leads to larger discrepancy values. Data from students of four different grade-years were evaluated and therefore a few items, only administered to students of grade one, are relatively easy with low discriminations. These items can still be used to assess the math skills of the grade-one students but they are not very useful for assessing the skills of the other students.

The grade-specific population parameters are given in Table 1. It can be seen that the average grade-means are increasing from grade one to grade four. The student performances significantly improve from year to year. The mean and variance of

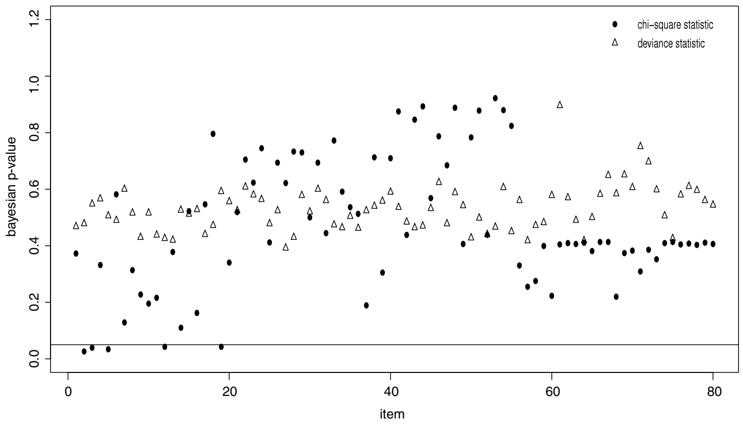

Fig. 4. Posterior predictive *p*-values corresponding to the item-based discrepancy measure based on Pearson (chi-square statistic) and deviance residuals (deviance statistic).

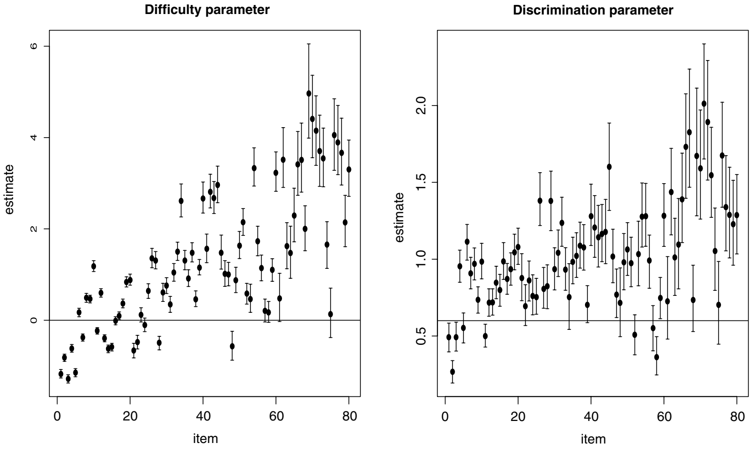

Fig. 5. Difficulty and discrimination parameter estimates and 95% credible intervals.

Table 1
Population parameter estimates with 95% credible intervals.

	Grade mean			Grade variance		
Grade	Mean	SD	CI 95%	Mean	SD	CI 95%
(Reference) 1	0	–	–	1	–	–
2	1.06	0.05	[0.97, 1.17]	0.77	0.07	[0.64, 0.92]
3	1.44	0.06	[1.33, 1.56]	0.39	0.04	[0.31, 0.48]
4	2.13	0.09	[1.97, 2.30]	0.59	0.08	[0.44, 0.77]

the first grade were fixed to zero and one, respectively, to identify the scale. The distribution of the student scores in the third and fourth year have a smaller standard deviation compared to the first two years, which means that the student performances are more alike in the last two years. After two years of education, the students are more alike with respect to their math performances. The standard deviations of the estimates are not equal across grades since the number of students differ over grades.

In Fig. 5, the posterior mean estimates of difficulty and discrimination parameters of all test items are given. The left subplot represents the difficulty estimates of the test items. The students in grade one responded to the first twenty items, which are also the easiest test items. The corresponding credible intervals are narrow since response data of subjects in grade two were also used for estimating the mean item difficulties. The estimated difficulties of items 20–40 are higher. These test items were made by students in grade two and three. The test items 40–60 are made by students in grade three and four and these items show more variation in item difficulty. Some of these items appear to be too easy, since they are of comparable difficulty with items administered to grade one students. The last twenty items were only administered to students in grade four, which are the most difficult items. They show that the students of grade four need significantly more math skills than students from earlier grades to accomplish the items. The estimated item discriminations are represented

Table 2
Pearson's between-grade correlations between latent scores and p-values.

	Grade 2	Grade 3	Grade 4
Grade 1	0.109 (0.072)	0.045 (0.456)	−0.027 (0.651)
Grade 2	–	0.015 (0.798)	0.215 (0.001)
Grade 3		–	0.083 (0.170)

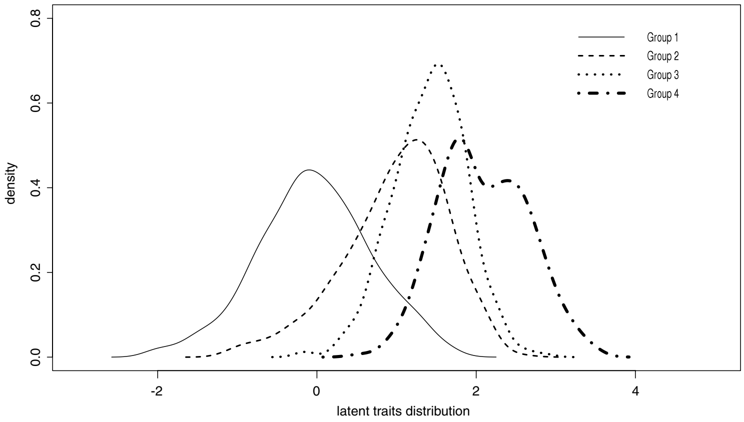

Fig. 6. Latent score distributions per grade year on a common scale.

in the right subplot. They also show an increasing pattern over items, where the more difficult items also discriminate more. This is a common relationship between item characteristics.

Finally, the between-grade independency was investigated, since it was assumed that students are nested within grade years but are assumed to be independent over grade years. Therefore, the between-grade correlations between latent math scores were examined using Pearson's correlation. In each MCMC iteration, Pearson's (between-grade) correlations were computed and the averages over iterations were considered to be an estimate of the marginal posterior Pearson's correlations. The estimates and corresponding p-values, which refer to the hypothesis of a non-zero correlation, are given in Table 2. Although some correlation was expected between two consecutive grade-years, it turned out that there is one significant correlation between the math scores of grade two and four. This significant correlation is considered to be a coincidence and is therefore not parameterized in the MGM to avoid any overfitting of the data. The other non-significant correlations support the assumption that the scores are not correlated over grade-years.

6.2. A Bayesian grade-comparison analysis

The simultaneous calibration of students from different grade-levels makes it possible to compare between-grade student performances. High and low-performing students are most often detected with respect to the mean grade level. In the present joint calibration, student achievements can also be compared to the mean average of the other grades. To identify or low- and high-performing students, it is of interest to compare their skills to the skills of students from consecutive grades.

In Fig. 6, the grade-specific distributions of the latent sores are represented. It can be seen that the latent score distribution of grade one has an overlap with the score distributions of the other grades. This means that some students in grade one have math skills that are comparable with students from grade two, three, and four. High-performing students in grade one can outperform students from grade four. Note that individual posterior probability statements can be made under the model to evaluate the posterior probability that a grade-one student outperforms a student from another grade level. Due to the fact that all students are calibrated on one scale, it follows that

$$P\left(\theta_j > \theta_{j'} \mid \mathbf{y}, j \in \text{grade k}, j' \notin \text{k}\right) = \int\int I\left(\theta_j > \theta_{j'}\right) p\left(\theta_j, \theta_{j'} \mid \mathbf{y}\right) d\theta_j d\theta_{j'}$$
$$\approx M^{-1}\sum_{m=1}^{M} I\left(\theta_{jk}^{(m)} > \theta_{j'k'}^{(m)}\right),$$

where $(\theta_j^{(m)}, \theta_{j'}^{(m)})$ are draws from the joint posterior using the MCMC scheme and $k \neq k'$.

Without interest in any specific student, the 25% and 75% quantile values for each grade-year distribution were considered. In each grade year, the points correspond to students with math skills ranked at the 25th percentile (lower quartile) and 75th percentile (upper quartile). The posterior probability of a correct response to each each item was computed for students at the 25th and 75th percentile in each grade year.

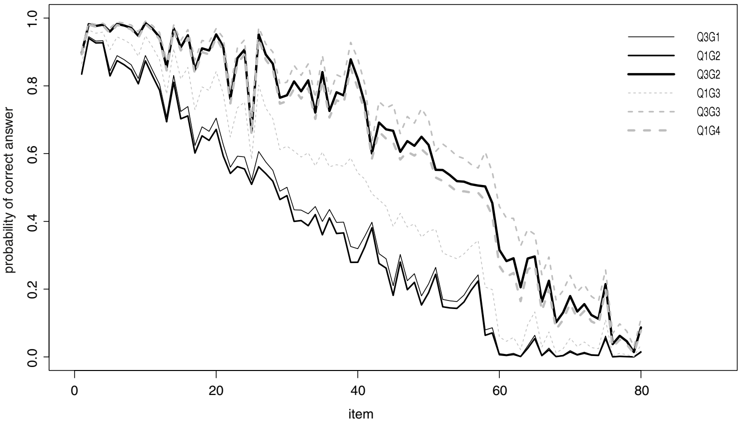

Fig. 7. Posterior probability of a correct response to items ordered by item difficulty for math achievements at the 25th (Q1)and 75th (Q3) percentile.

In Fig. 7, the so-called person-response curves for each 25th percentile, denoted as Q1, and the 75th percentile, denoted as Q3, are plotted for each grade year. The G1 to G4 refer to grade one to grade four. The person-response curve is a function that relates the probability of a correct response to the item difficulty. The person-response function is monotonically decreasing when all items have the same discrimination level. Here, the test items have different discrimination values but the item discriminations are invariant over grade years. Therefore, peaked curves are obtained but with common peaks across person-response curves.

The curve is often used to see whether persons are consistent in their response behavior, since it is to be expected that the more difficult items have lower success probabilities. Therefore, the slope of a person-response curve is steep for consistent response behavior and flat for inconsistent response behavior. In Fig. 7, the slopes of the person-response curves are similar and show consistent response behavior. The person-response curves of the 25th percentile of grade year two and three are flat for the last twenty items, since these items are too difficult. Then, there is no relationship between the item difficulties and the success probabilities.

By comparing the person-response curves across grade years, it can be seen how students from a grade year perform in comparison to the other grade years. It can be seen that the 75th percentile of grade one perform more or less the same as the 25th percentile of grade two. The best performing students of the first year already match the performances of the lowest performing students of the second year. It is remarkable that the 75th percentile of grade two already matches the performance of the 25th percentile of grade four. The students differ by two years of education but their performances are more or less the same. Note that the estimated latent scores of students in grade two are also shrunk towards the mean grade level, which is much lower than the mean level of grade four. Furthermore, the estimated latent scores in grade four are shrunk towards the mean level of grade four.

7. Comments and conclusions

A Bayesian multiple group modeling approach is proposed for handling response data from respondents clustered in groups. The developed Bayesian methods for the MGM include an estimation method based on MCMC and different posterior predictive assessment tools. In a simulation study, the MCMC algorithm showed a good recovery of the model parameters. It was further shown that the relative bias of the estimated model parameters was smaller than those of Bilog–MG. The assessment tools were shown to be useful in evaluating the fit of the MGM and model violations of various assumptions could be detected.

A generalized MGM framework has been discussed. The group-specific latent variable distribution might not always be normal, and it is shown that adjustment can be made to use a skewed latent variable distribution to model an asymmetric population distribution of the latent variable. This includes the skewed latent variable approach of Azevedo et al. (2011).

The MGM is of particular interest when latent scores are to be estimated on a common scale across groups. The identification of the MGM is based on a set of common (anchor) items that have common item characteristics across groups. When linking the group-specific scales with anchor items, the generalized MGM can be used to evaluate differential item functioning and to test whether items are measurement invariant across groups in a very flexible way.

Acknowledgments

The authors are thankful to CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) from Brazil, for the financial support through a Doctoral Sandwich Scholarship granted to the first author under the guidance of the two others. 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. Finally, the authors are thankful to Danylo Henrique Moya, for the support in the simulations.

Appendix

The model defined by Eqs. (5)–(9) can be identified, by following the approach of Fox and Glas (2001), due to the structure given by Eqs. (5) and (6). That is, by fixing the discrimination and difficulty parameters of any item, in 1 and 0, respectively. This is equivalent to eliminate one equation in (8) and another in (9).

The modified step concerning the Metropolis–Hastings within Gibbs sampling algorithm with independent prior distributions for the population parameters (scheme 2) is given by:

Simulate item parameters $(a_i, b_i)'$ from a proposal distribution

$$\begin{aligned}&a_i^{(*)} \mid a_i^{(t-1)} \sim \text{log-normal}(\ln a_i^{(t-1)}, \sigma_a^2)\\&b_i^{(*)} \mid b_i^{(t-1)} \sim N(b_i^{(t-1)}, \sigma_b^2)\end{aligned}$$

and accept $\boldsymbol{\zeta}_i^{(t)} = \boldsymbol{\zeta}_i^{(*)}$ with probability:

$$\pi_i\left(\boldsymbol{\zeta}_i^{(t-1)}, \boldsymbol{\zeta}_i^{(*)}\right) = \min\left\{R_{\zeta_i}, 1\right\},$$

where

$$R_{\zeta_i} = \frac{L(\boldsymbol{\theta}_{\dots}^{(t)}), \boldsymbol{\zeta}_i^{(*)}}{L(\boldsymbol{\theta}_{\dots}^{(t)}), \boldsymbol{\zeta}_i^{(t-1)}} \times \frac{\frac{1}{\left(a_i^{(*)}\right)^2} \exp\left\{-\frac{1}{2\psi_a}\left(\ln a_i^{(*)} - \mu_a\right)^2\right\}}{\frac{1}{\left(a_i^{(t-1)}\right)^2} \exp\left\{-\frac{1}{2\psi_a}\left(\ln a_i^{(t-1)} - \mu_a\right)^2\right\}} \times \frac{\exp\left\{-\frac{1}{2\psi_b}\left(b_i^{(*)} - \mu_b\right)^2\right\}}{\exp\left\{-\frac{1}{2\psi_b}\left(b_i^{(t-1)} - \mu_b\right)^2\right\}},$$

otherwise, set $\boldsymbol{\zeta}_i^{(t)} = \boldsymbol{\zeta}_i^{(t-1)}$.

The convergence of the algorithms of scheme 1 can be speeded up by using the approach of Gonzalez (2004). Then, for $i = 1, \dots, I$, simulate item parameters $(a_i, b_i)'$ as in step 4 of MCMC scheme 1. Furthermore,

- Fix the simulated values by doing $\overline{\boldsymbol{\zeta}}_i = (b_i, a_i/b_i) = (b_i, \overline{a}_i)$.
- Generate a random scale, say, $\nu \sim f(\nu)$ (e.g. from a gamma distribution) and

  Fix $b_i^{(t)} = \nu b_i$, $a_i^{(t)} = \nu b_i \overline{a}_i$ with probability

$$\min\left\{\frac{L(b_i^{(t)}, a_i^{(t)}\boldsymbol{y}_{i.})\pi(b_i^{(t)}, a_i^{(t)})}{L(b_i, a_i\boldsymbol{y}_{i.})\pi(b_i, a_i)}\frac{f(1/\nu)}{f(\nu)}|\nu^2|, 1\right\}$$

  where $\pi(\cdot)$ is the prior distribution of $\boldsymbol{\zeta}_i$, otherwise set $b_i^{(t)} = b_i$, $a_i^{(t)} = b_i\overline{a}_i$.

Another alternative to speed up the convergence is to simulate simultaneously augmented data and latent traits. Therefore, modify step 1 as follows:

For $j = 1, \dots, n_k$ $k = 1, \dots, K$, simulate $\boldsymbol{Z}_{\cdot jk}|(\cdot)$, marginalized over the latent traits, that is, simulate $\boldsymbol{Z} \sim N_I(\boldsymbol{a}\mu_{\theta_k} - \boldsymbol{b}, \psi_{\theta_k}\left[\boldsymbol{I}_I + \boldsymbol{a}\boldsymbol{a}'\right])\mathbb{I}_A(\boldsymbol{z} \cdot jk)$, where A corresponds to the parameter space of each component of $\boldsymbol{Z}_{\cdot jk}$. This simulation can be easily implemented by using the algorithm proposed by Geweke (1991), for example.

The description of the modified steps concerning the other algorithms are available upon request from the authors.

A.1. Computation of posterior predictive P-values

For the calculations of the p-values related to the proposed discrepancy measures, item responses are generated through the current simulated values of all model parameters (latent traits, item, and population parameters). The parameter-dependent discrepancy measures are evaluated using posterior predictive data and the MCMC samples of the parameters. A procedure for the computation of the expected count scores is given by Béguin and Glas (2001). The expected count scores $E(n_{l,k})$ can also be computed using the conditional item response probabilities given the latent traits. Then, the conditional probability is computed of each response response pattern which leads to a score of $n_{l,k}$, and let $P^{(t)}$ denote the sum of corresponding conditional probabilities in iteration t. Then, the conditional expected score can be calculated from $E(n_{l,k}) = NP^{(t)}$. Note that this conditional expected score will lead to a parameter-dependent discrepancy measure. The variance can be calculated as $V(n_{l,k}) = NP^{(t)}\left(1 - P^{(t)}\right)$, where t indicates that the probability is calculated by using the simulated values of the parameters in iteration t.

References

Albert, J., 1992. Bayesian estimation of normal ogive item response curves using Gibbs sampling. Journal of Educational Statistics 17, 251–269.

Andrade, D.F., Tavares, H.R., Valle, R.C., 2000. Item Response Theory: Concepts and Applications. 14^0 SINAPE, ABE (in Portuguese).

Azevedo, C.L.N., Andrade, D.F., 2010. An estimation method for latent traits and population parameter for the nominal response model. Brazilian Journal of Probability and Statistics 24, 415–433.

Azevedo, C.L.N., Bolfarine, H., Andrade, D.F., 2011. Bayesian inference for a skew-normal IRT model under the centred parameterization. Computational Statistical & Data Analysis 55 (1), 353–365.

Béguin, A.A.B., Glas, C.A.W., 2001. MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika 66 (4), 541–561.

Bock, R.D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. Psychometrika 46, 433–459.

Bock, D.R., Zimowski, M.F., 1997. The multiple groups IRT. In: van der Linden, Wim J., Hambleton, Ronald K. (Eds.), Handbook of Modern Item Response Theory. Springer-Verlag.

Chen, M.-H., Shao, Q.-M., Ibrahim, J.G., 2000. Monte Carlo methods for Bayesian computation, first ed. Springer-Verlag, New York.

da-Silva, C.Q., Gomes, A.E., 2011. Bayesian inference for an item response model for modeling test anxiety. Computational Statistica & Data Analysis 55, 3165–23182.

De Ayala, R.J, Sava-Bolesta, M., 1999. Item parameter recovery for the nominal response model. Applied Psychological Measurement 23 (1), 3–19.

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

Fox, J.-P., 2010. Bayesian Item Response Modeling: Theory and Applications, first ed. Springer-Verlag, New York.

Fox, J.-P., Glas, C.A.W., 2001. Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika 66, 269–286.

Fox, J.-P., Glas, C.A.W., 2005. Bayesian modification indices for IRT models. Statistica Neerlendica 59, 95–106.

Gamerman, D., Lopes, H.F., 2006. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, second ed. Chapman & Hall, London.

Geweke, J., 1991. Efficient simulation from the multivariate normal and Student-*t* distributions subject to linear constraints and the evaluation of constraint probabilities, Technical report.

Glas, C.A.W., 1998. Detection of differential item functioning using Lagrange multiplier tests. Statistica Sinica 8, 647–667.

Gonçalves, F.B., 2006. Bayesian analysis of item response theory: a generalized approach. Master Thesis. in Portuguese, Department of Statistics, Federal University of Rio de Janeiro (unpublished).

Gonzalez, R.L., 2004. Data augmentation in the Bayesian multivariate probit model. In: Sheffield Economic Research Paper Series.

Kim, Seock-Ho, Cohen, A.S., Park, Tae-Hak, 2005. Detection of differential item functioning in multiple groups. Journal of Educational Measurement 32 (3), 261–276.

Kolen, M.J., Brennan, R.L., 2004. Test Equating — Methods and Pratices, second ed. Springer-Verlag, New York.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman & Hall, London.

Mislevy, R.J., 1984. Estimating latent distributions. Psychometrika 49, 359–381.

Mislevy, R.J., 1986. Bayes modal estimation in item response models. Psychometrika 51, 177–195.

Montgomery, D.C., 2004. Design and Analysis of Experiments, sixth ed. Chapman & Hall, London.

Múthen, B., Lehman, J., 1985. Multiple group IRT modeling: applications to item bias analysis. Journal of Educational and Behavioral Statistics 10 (2), 365–376.

Penfield, R.D., 2001. Assessing differential item functioning among multiple groups: a comparison of three Mantel–Haenszel procedures, 14, 3, 235-259.

Sinharay, S., 2006. A Bayesian item fit analysis for unidimensional item response theory models. British Journal of Mathematical and Statistical Psychology 59, 429–449.

Sinharay, S., Johnson, M.S., Stern, H., 2006. Posterior predictive assessment of item response theory models. Applied Psychological Measurement 30 (4), 298–321.

Soares, T.M., Gonçalves, F.B., Gamerman, D., 2009. An integrated Bayesian model for DIF analysis. Journal of Educational and Behavourial Statistics 34 (3), 348–377.

Stern, H.S., Sinharay, S., 2005. Bayesian model checking and Model Diagnostics. In: Dey, Dipak K., Rao, C.R. (Eds.), Bayesian Modelling, Thinking and Computation. In: Handbook of Statistics, vol. 25. Elsevier, The Netherlands.

Zimowski, M.F., Muraki, E., Mislevy, R.J., Bock, R.D., 1996. Bilog–MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items. SSI: Scientific Software, Chicago.