

# Redefining Item Response Models for Small Samples

Jean-Paul Fox 

University of Twente

*Popular item response theory (IRT) models are considered complex, mainly due to the inclusion of a random factor variable (latent variable). The random factor variable represents the incidental parameter problem since the number of parameters increases when including data of new persons. Therefore, IRT models require a specific estimation method and large samples for accurate parameter estimation. The two-parameter IRT model is redefined by analytically integrating out the random person factor in the latent response formulation of the model to make it suitable for small sample applications. This IRT Bayesian covariance structure model (IRT-BCSM) describes the clustering of (latent) responses by persons through a structured covariance matrix in which a common covariance parameter represents the dependence implied by a unidimensional random factor variable. The IRT-BCSM is a latent-variable-free model and consists of (fixed) item parameters and a common covariance parameter, where person parameters can be post-hoc sampled. An efficient Gibbs sampler is proposed for parameter estimation. In simulation studies, the performance of the IRT-BCSM is compared to two-parameter IRT models for small samples, and results show an optimal performance of the IRT-BCSM even in sample sizes as small as 50 to 100 persons and 5 to 10 items. Generalizations of the IRT-BCSM show that a redefinition of more complex IRT models also lead to much more efficient parameterizations, which can broaden the scope of IRT applications.*

Keywords: *Bayesian; covariance structure modeling; item response theory; small samples*

## 1. Introduction

Traditional item response theory (IRT) models are built for a two-factor design. There is one factor that describes the clustering of item responses by persons (person factor), and the second factor describes the clustering of item responses by items (item factor). The required sample size for accurate parameter estimation is determined by the number of items and the number of persons. Requiring a sufficient number of items and persons is typical for the cross-classified two-factor design of an IRT model. It is well known that increasing the

number of persons (items) improves the accuracy of the item (person) parameter estimates (e.g., Reckase, 2009; Reise & Yu, 1990).

Determining the required sample size for an IRT model is complicated. Item response observations are correlated due to the two-factor design. Therefore, a required sample size depends on the fit of an IRT model. When both factors are well supported by the data—the corresponding eigenvalues explain a sufficient amount of variance in the observed item responses—the data are (relatively) informative and the required sample size will be lower than when one or both factors are poorly supported by the data. This dependence of the required sample size on the fit of an IRT model leads to estimation problems of IRT parameters with small samples. The smaller the sample the more likely it is that one of the factors is not well supported by the data, which leads to estimation problems. This principle follows for instance from (exploratory) factor analysis, in which latent variables are selected with eigenvalues greater than one (Kaiser–Guttman rule), since they explain the relevant covariance among observed variables (Brown, 2015).

The required sample size for IRT models has been thoroughly investigated. For dichotomously scored items, it is often suggested that at least 500 persons are needed for accurate item parameter estimation (De Ayala, 2009; Stone, 1992). However, various studies examining the minimum required sample size has led to different recommendations. König et al. (2020) exploited a hierarchical prior structure for the item parameters and pooled data information across items and obtained accurate estimates for samples with only 100 persons. However, the borrowing-strength principle only works when having a sufficient number of items and persons. For instance, in mixed effects models, the number of grouping levels can greatly influence the accuracy of the (random effect) variance estimates (Gelman & Hill, 2006, p. 247). Shrinkage is strongest for groups with small sample sizes, where data from other groups are used to improve the precision of the group-specific estimates. However, particularly for small number of groups, this partial pooling of information can lead to poor group-level estimates, where the within-group and between-group data information is not optimally weighted due to inaccurate random effect variance estimates (Bryk & Raudenbush, 1992).

Furthermore, the accuracy of the estimation depends on the amount of shrinkage, which depends on the variation in item and person parameters and the sensitivity of the priors in inducing shrinkage. Specifically for small samples, the appropriate control of the level of shrinkage in the parameter estimates requires a priori knowledge about the (factor) variances of the person and item parameters, which are usually unknown. Harwell and Janosky (1991) studied the impact of the prior variances for the discrimination and difficulty parameters for different sample sizes on their posterior distributions. They showed that the estimation results were inaccurate for a small number of persons (250 or fewer) and for larger prior variances. It was concluded that the specification of the prior

variances for a small number of items and persons is really important for accurate estimation, but that guidelines are needed to select prior variances.

The role of the random factor in the sample size requirements is connected to its role in the construction of estimation methods, in which it leads to dealing with the *incidental parameter problem* (Kalbfleisch & Sprott, 1970; Lancaster, 2000). The problem is characterized by the fact that, given a fixed number of items, new observations lead to new (person) parameters. To deal with the problem, a conditional and an unconditional approach can be taken. For the two-parameter model, there is no sufficient statistic to condition on to eliminate the person parameter from the likelihood. Furthermore, for a fixed number of items, joint maximum likelihood (ML) estimation can lead to bias in the item parameter estimates, since standard asymptotic theory for ML estimation does not apply. In the unconditional approach, a (prior) distribution for the person parameters is defined and an integrated or marginal likelihood function is obtained by integrating over the person parameters. The maximization of this integrated likelihood function will give consistent ML estimators for the item parameters. The estimation procedure is known as marginal maximum likelihood (MML) estimation (Bock & Aitkin, 1981). It is common to assume a normal distribution for the person parameters, and the marginal likelihood is obtained through integration using Gauss–Hermite quadrature.

When applying Gauss–Hermite quadrature to compute the marginal likelihood, the structure of the IRT model remains unaltered. In the conditional likelihood, the person parameters are replaced by quadrature nodes, and the marginal likelihood for the discrimination and difficulty parameters is approximated by a weighted sum over the conditional likelihood. Therefore, obtaining stable parameter estimates depends on the frequency distribution of observed response patterns in relation to the item score patterns. A serious discrepancy between the expected number of correct responses and the observed number of correct responses for each quadrature node—each node represents a specific response pattern—can lead to unstable item parameter estimates, which is likely to occur with small samples. The likelihood equations for the item parameters under MML estimation show this dependence on the number of persons responding in each response pattern and the expected frequency under the two-parameter (Probit) IRT model (Bock & Aitkin, 1981). The discrimination parameter estimate is more sensitive to such discrepancies than the difficulty parameter estimate, since it is based on a linear relationship between expected item scores and quadrature nodes (representing different response patterns). Therefore, MML estimation does not relax sample size restrictions for the IRT model.

Another unconditional estimation method is Markov chain Monte Carlo (MCMC). Although priors for the item parameters can improve the parameter estimation, the posterior sampling of item parameters given sampled person parameters implies sample size restrictions on the number of persons and items.

Stable posterior draws of the item parameters can be obtained, leading to accurate estimates of posterior quantities, but it requires stable posterior draws of the person parameters. Thus, in the same way as for MML, MCMC for IRT parameter estimation places restrictions on the minimum number of items and persons to obtain accurate parameter estimates.

To obtain accurate item parameter estimates for small samples, the IRT model is redefined to avoid the incidental parameter problem. A straightforward approach is taken by analytically integrating out the person parameters. The resulting IRT model, after integrating out the person parameters, only contains structural parameters (Kalbfleisch & Sprott, 1970). The analytical integration is easy when using a latent response variable formulation of the IRT model (Fox, 2010). This redefined IRT model is referred to as the IRT Bayesian covariance structure model (IRT-BCSM). The IRT-BCSM generalizes former BCSM approaches to Bayesian item response modeling by including discrimination parameters. Fox et al. (2017) considered the one-parameter IRT model and discussed the covariance structure model implied by a random person factor. Fox et al. (2020) considered the implied covariance structure of random testlet effects without including discrimination parameters in the covariance structure. In contrast to these approaches, the proposed IRT-BCSM considers a heterogeneous covariance structure for the latent response patterns, in which the discrimination parameters can modify the homogeneous dependence among item responses of the same response pattern. The IRT-BCSM will relax the sample size restrictions for item parameter estimation, since it does no longer depend on any person parameters or numerical integration over the distribution of person parameters.

After introducing the (Probit) two-parameter IRT model, the sample size restrictions for accurate parameter estimation are discussed. Then, the IRT-BCSM is introduced, in which the structured covariance matrix represents the dependence implied by a random person factor, while accounting for the discrimination effects. The derivation of the posterior distributions of the discrimination and common covariance parameters are discussed in detail since they contain novel elements in the posterior computation of components of the structured covariance matrix. In the Bayesian approach, priors are defined for the item parameters and the common covariance parameter. An efficient Gibbs sampling algorithm is proposed for parameter estimation. Then, simulation results for several small sample test designs are presented, which include a comparison in performance between the well-known two-parameter IRT model and the IRT-BCSM. In the final section, conclusions and IRT-BCSM generalizations are discussed.

## 2. Redefining IRT Models: The IRT-BCSM

For the two-parameter (normal ogive) IRT model, the probability of a correct response,  $Y_{ik}$ , of person  $i$  ( $i = 1, \dots, n$ ) to item  $k$  ( $k = 1, \dots, m$ ) is given by

$$\begin{aligned}
 P(Y_{ik} = 1 | \theta_i, a_k, b_k) &= \Phi(a_k \theta_i - b_k) \\
 \theta_i &\sim \mathcal{N}(0, \tau),
 \end{aligned}
 \tag{1}$$

where function  $\Phi(\cdot)$  represents the normal cumulative distribution function. The  $\theta_i$  is the random factor person,  $a_k$  the discrimination parameter, and  $b_k$  the difficulty parameter. The parameter  $\tau$  is the random factor variance representing the level of variance in the population between persons in their trait levels (random factor levels). In the Bayesian two-parameter IRT model (Fox, 2010), priors are defined for the item parameters and the factor variance and possibly hyper priors for the prior parameters.

The Probit IRT model in Equation 1 allows introducing normally distributed latent variables, which determine the discrete item observations through a threshold specification (Albert & Chib, 1993; Fox, 2010). Therefore, assume a normally distributed random variable  $Z_{ik}$  with mean  $a_k \theta_i - b_k$  and variance one and truncated to be greater than zero if the item response is one and less than zero otherwise. Then, in this latent response formulation, the random factor can be easily integrated out (Fox et al., 2017, 2020). It follows that the latent responses of person  $i$  are multivariate normally distributed:

$$\begin{aligned}
 \mathbf{Z}_i &= -\mathbf{b} + \mathbf{E}_i \\
 \mathbf{E}_i &\sim \mathcal{N}(0, \Sigma) \\
 \Sigma &= \tau \mathbf{a} \mathbf{a}^t + \mathbf{I}_m.
 \end{aligned}
 \tag{2}$$

Without conditioning on the random factor person, the item responses of each person  $i$  are correlated with a common covariance  $\tau$ , which resembles for  $\tau > 0$  the random factor variance in the two-parameter IRT model in Equation 1. The discrimination parameters  $\mathbf{a} = (a_1, \dots, a_m)^t$  modify the common covariance among (latent) item responses. Thus, the first component of the covariance matrix,  $\tau \mathbf{a} \mathbf{a}^t$ , represents the dependence among item responses of person  $i$ —this is the implied dependence by the random factor person modified by the discriminations  $\mathbf{a}$ . The other component in the covariance matrix,  $\mathbf{I}_m$ , represents independence across items with a common error variance of one (for reasons of identification).

The revised IRT model in Equation 2 is referred to as an IRT-BCSM. The random factor person is integrated out, and the dependence among item responses is modeled though a structured covariance matrix. The mean component of the model represents the item difficulties, and the covariance matrix contains the component describing the dependence among item responses implied by a random person factor. Therefore, the IRT-BCSM in Equation 2 still represents the dependence structure of a two-factor (IRT) model but without using a random person factor. The discrimination parameters are also no longer slope parameters, which modify item-specific contributions of a person effect (ability) on a correct

response. Instead the item discriminations describe the item-specific covariances of (latent) responses clustered by persons.

The IRT-BCSM represented in terms of the observed data leads to a multivariate Probit model, in which the probability of observing response pattern  $\mathbf{y}_i$  is given by

$$P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{a}, \mathbf{b}, \tau) = \int_{\Omega(y_{i1})} \dots \int_{\Omega(y_{im})} \Phi_m(\mathbf{z}_i | \mathbf{b}, \Sigma) d\mathbf{z}_i,$$

where

$$\Omega(\mathbf{y}_i) = \{\mathbf{z}_i : z_{ik} \leq 0 \text{ if } y_{ik} = 0, z_{ik} \geq 0 \text{ if } y_{ik} = 1\}.$$

It follows that the parametrization of the IRT-BCSM is more efficient than that of a two-parameter IRT model, since the IRT-BCSM does not contain any incidental parameters. For a fixed number of items, the number of model parameters does not change when observing new data. The mean structure of the IRT-BCSM represents a fixed factor (item) design in contrast to the cross-classified two-factor (person and item) design of the two-parameter IRT model. Furthermore, the dependence structure implied by the clustering of responses by persons is modeled by a single covariance parameter  $\tau$ . This makes the redefined IRT model a latent-variable-free IRT model. The latent variable, referred to as the person random factor, includes the assumption of normality and is known to increase the sample size requirements for accurate estimation.

The IRT-BCSM also requires identification restrictions. When introducing a general mean across items, the sum of the item difficulty parameters needs to be constrained. A constraint is needed for the item discrimination parameters to identify the  $\tau$  parameter. Furthermore,  $\tau$  is allowed to be negative in the IRT-BCSM, since it is a covariance parameter with the restriction that the covariance matrix needs to be positive definite. By allowing  $\tau$  to be negative, the item discriminations are restricted to be positive. The constraints and prior distributions for the model parameters are discussed in Section 3 and Appendix A.

### 3. Posterior Computation

Posterior computation is performed using a Gibbs sampling algorithm. The algorithm consists of four main sampling steps: sampling of latent response variables, item difficulty and discrimination parameters, and a common covariance parameter. The derivation of the posterior distribution for the discrimination parameters follows from deriving the conditional distribution from the multivariate normal distribution for the (latent) item response data. The posterior distribution of the covariance parameter is based on an eigendecomposition of the covariance matrix. These parts represent the novel elements of the posterior computations.

The posterior distributions of the remaining parameters follow in a straightforward manner, which are discussed in Appendix A.

### 3.1 Posterior Distribution Item Parameters

To derive the posterior distributions of the item parameters, conditional distributions of the multivariate normal distribution in Equation 2 are considered using an analytical expression for the covariance matrix  $\Sigma$ . Through the analytical form of the inverse of this covariance matrix, the mean and variance of the conditional distributions can be explicitly defined.

Using the Sherman–Morrison formula, the inverse of the covariance matrix  $\Sigma$  can be expressed as

$$\begin{aligned} \Sigma^{-1} &= \mathbf{I}_m - \frac{\boldsymbol{\tau}\mathbf{a}\mathbf{a}^t}{\mathbf{a}^t\mathbf{a}\boldsymbol{\tau} + 1} \\ &= (\mathbf{I}_m - \mathbf{A}/\lambda)\mathbf{T}^{-1}, \end{aligned}$$

where  $\lambda = \mathbf{a}^t\mathbf{a} + 1/\boldsymbol{\tau}$ ,  $\mathbf{A} = \mathbf{a}\mathbf{a}^t$  but  $\text{diag}(\mathbf{A}) = 0$ , and  $\mathbf{T} = \text{diag}(\sigma_1^2, \dots, \sigma_m^2)$  with  $\sigma_k^2 = a_k^2\boldsymbol{\tau} + 1$  representing the error variance of item  $k$ .

Consider the distribution of  $\mathbf{Z}_k$  to derive the posterior distribution for  $\sigma_k^2$  and  $b_k$ ,

$$p(\mathbf{z}_k | \sigma_k^2, b_k) = (2\pi\sigma_k^2)^{-n/2} \exp\left(\frac{-1}{2\sigma_k^2} \sum_i (z_{ik} + b_k)^2\right). \quad (3)$$

The posterior distribution for  $\sigma_k^2$  and  $b_k$  follows in a straightforward manner and is represented in Appendix A.

The discrimination parameters are covariance parameters in the IRT-BCSM in Equation 2, and a normal prior is defined with mean  $\mu_a$  and standard deviation  $\sigma_a$ . These prior parameters have hyper priors (see Appendix A). Consider the conditional distribution of  $Z_{ik}$  given  $\mathbf{Z}_{i,-k}$ , where  $\mathbf{Z}_{i,-k}$  equals  $\mathbf{Z}_i \setminus \{Z_{ik}\}$ . This is a conditional normal distribution for which the mean and variance can be derived from standard properties of the multivariate normal distribution and using the analytical form of the inverse of the covariance matrix. It follows that

$$z_{ik} | \mathbf{z}_{i,-k}, \mathbf{a}, \mathbf{b}, \boldsymbol{\tau} \sim \mathcal{N}\left(-b_k + \frac{a_k}{\lambda_{-k}} \sum_{l \neq k} a_l (z_{il} + b_l), \sigma_z^2\right), \quad (4)$$

where  $\sigma_z^2 = 1 + a_k^2/\lambda_{-k}$  with  $\lambda_{-k} = \mathbf{a}_{-k}^t\mathbf{a}_{-k} + 1/\boldsymbol{\tau}$  and  $\mathbf{a}_{-k} = \mathbf{a} \setminus \{a_k\}$ . A posterior sample of  $\sigma_k^2$  can be used to obtain a draw for  $\sigma_z^2$ , since  $(\sigma_k^2 - 1)/(\boldsymbol{\tau}\lambda_{-k}) = \sigma_z^2 - 1$ .

The conditional distribution of the  $Z_{ik}$  is used to derive the posterior distribution of  $a_k$ . Let  $x_{ik} = \mathbf{a}'_{-k}(\mathbf{z}_{i(-k)} + \mathbf{b}_{-k})/\lambda_{-k}$ . Then, the posterior distribution of  $a_k$  is normal:

$$a_k | \mathbf{z}_k, \mathbf{x}_k, b_k, \mu_a, \sigma_z^2, \sigma_a^2 \sim \mathcal{N}(\omega_k(\mathbf{x}'_k(\mathbf{z}_k + b_k)/\sigma_z^2 + \mu_a/\sigma_a^2), \omega_k), \quad (5)$$

where  $\omega_k = (\mathbf{x}'_k \mathbf{x}_k / \sigma_z^2 + 1 / \sigma_a^2)^{-1}$ . The hyper prior for  $\sigma_a^2$  is an inverse-gamma prior, but this hyper prior can be adjusted to reduce shrinkage effects. In the simulation studies, it is shown that the inverse-gamma prior for  $\sigma_a^2$  leads to accurate estimation results. The posterior distribution of the discrimination parameter can be truncated from below to restrict the domain to positive values.

The latent response data  $Z_{ik}$  are sampled using their conditional distribution, Equation 4, but truncated to the domain  $Z_{ik} \leq 0$  if  $Y_{ik} = 0$  or  $Z_{ik} > 0$  if  $Y_{ik} = 1$ .

### 3.2 Posterior Distribution Common Covariance Parameter

The posterior distribution of the common covariance  $\tau$  is retrieved through an eigenvalue decomposition of the (symmetric and positive definite) covariance matrix  $\Sigma$ . The component  $\mathbf{a}\mathbf{a}'$  has rank 1 and, therefore, only one nonzero eigenvalue. Therefore, one eigenvalue of  $\Sigma$  equals  $\mathbf{a}'\mathbf{a}\tau + 1$  and  $m - 1$  eigenvalues equal one. Let  $\mathbf{D}$  represent the eigenvectors corresponding to the  $m$  eigenvalues. Then, the distribution of  $\mathbf{D}(\mathbf{Z}_i + \mathbf{b})$  is normal with mean zero and covariance matrix

$$\begin{aligned} \mathbf{D}\Sigma\mathbf{D}' &= \mathbf{D}(\mathbf{a}\mathbf{a}'\tau + \mathbf{I}_m)\mathbf{D}' \\ &= \mathbf{D}\mathbf{a}\mathbf{a}'\mathbf{D}'\tau + \mathbf{I}_m \\ &= \mathbf{a}'\mathbf{a}\tau\mathbf{K}_m + \mathbf{I}_m, \end{aligned}$$

with  $\mathbf{K}_m$  a single-entry precision matrix with a one at position (1, 1) and all other elements zero. By definition of the eigenvalue decomposition, the covariance matrix of the response vector  $\mathbf{D}\mathbf{Z}_i$  is diagonal with the positive eigenvalues on the diagonal. There are  $m - 1$  eigenvalues of one and the other eigenvalue equals  $\tilde{\lambda} = \mathbf{a}'\mathbf{a}\tau + 1$ . The eigenvalue  $\tilde{\lambda}$  is restricted to be positive, which places a restriction on the parameter space of  $\tau$ . Therefore, a (conjugate) shifted inverse-gamma prior for  $\tau$  is defined (see Appendix C for the shifted inverse-gamma distribution), which includes a positive-definite restriction of the covariance matrix—restricting the  $\tilde{\lambda}$  to be greater than zero:

$$p(\tau|\mathbf{a}) \propto (\tau + 1/(\mathbf{a}'\mathbf{a}))^{-g_1-1} \exp\left(-\frac{g_2/(\mathbf{a}'\mathbf{a})}{\tau + 1/(\mathbf{a}'\mathbf{a})}\right), \quad (6)$$

where  $1/(\mathbf{a}'\mathbf{a})$  is the shift parameter. Then, the posterior distribution for  $\tau$  is also a shifted inverse-gamma distribution

$$p(\tau|\mathbf{z}, \mathbf{b}, \mathbf{a}) \propto (\tau + 1/(\mathbf{a}'\mathbf{a}))^{-g_1 - n/2 - 1} \exp\left(-\frac{\left(g_2 + \sum_i SS_i\right)/(\mathbf{a}'\mathbf{a})}{\tau + 1/(\mathbf{a}'\mathbf{a})}\right), \quad (7)$$

where  $SS_i = (\mathbf{D}'_1(\mathbf{z}_i + \mathbf{b}))^2$  and  $\mathbf{D}_1$  is the eigenvector corresponding to eigenvalue  $\tilde{\lambda}$ . It follows that the positive-definite restriction implies that  $\tau > -1/(\mathbf{a}'\mathbf{a})$ .

### 3.3 Post-Hoc Sampling of Person Parameters

The dependence implied by a structured covariance matrix with  $\tau > 0$ , Equation 2, can also be described by a random factor variable  $\theta_i$ , when implying a normal prior with mean zero and variance  $\tau$ . Subsequently, the conditional distribution of  $\mathbf{Z}_i$  given  $\theta_i$  is normal with mean  $(\mathbf{a}\theta_i - \mathbf{b})$ . Then, the posterior distribution of  $\theta_i$  is normal with mean and variance (Albert, 1992):

$$E(\theta_i|\mathbf{z}_i, \mathbf{a}, \mathbf{b}, \tau) = (\mathbf{a}'\mathbf{a} + 1/\tau)^{-1}(\mathbf{a}'(\mathbf{z}_i + \mathbf{b}))$$

$$V(\theta_i|\mathbf{z}_i, \mathbf{a}, \mathbf{b}, \tau) = (\mathbf{a}'\mathbf{a} + 1/\tau)^{-1},$$

respectively. Although  $\theta_i$  is not a model parameter, posterior draws for  $\theta_i$  can be obtained using MCMC draws for the item parameters and common covariance parameter  $\tau$ .

## 4. Simulation Study

Two simulation studies were performed. In the first simulation study, the performance of the IRT-BCSM in recovering the item parameters was compared to that of the two-parameter logistic regression (2PL) model and the two-parameter normal ogive (2PNO) model, while using a suitable estimation method for each model. In the second simulation study, the recovery of the parameters of the IRT-BCSM was examined for different small sample sizes.

### 4.1 Study 1: Model Comparison

A comparison was made between the performance of the IRT-BCSM, the 2PL model, and the 2PNO model for different sample sizes. For each model, a specific estimation method was used. For the 2PL model, the R-package *mirt* was used, which used an MML estimation method through an EM algorithm (Bock & Aitkin, 1981). The standard errors were computed through Fisher's information matrix, where Louis method (Louis, 1982) was used for computation of the observed information matrix. The 2PNO model was fitted using MCMC according to the Gibbs sampling algorithm of Albert and Chib (1993) and Johnson and Albert (1999), using non-informative priors for the item parameters. Therefore,

the R-package *sirt* (function *mcmc.2pno*) was used to estimate the model parameters, in which the latent variable variance was fixed to one to identify the model. A total of 6,000 MCMC iterations was made and the first 1,000 were the burn-in iterations. The posterior mean was used for parameter estimation and the posterior standard deviation for estimation precision.

For each data set, with sample sizes of  $n$  (50, 100, 500, 1,000) and for  $m = 10$ , discrimination and difficulty parameters were simulated from a normal distribution with mean 0.80 and standard deviation 0.10, and a mean of 0 and a standard deviation of 0.50, respectively. The average discrimination was set below one, to avoid sampling a high level of discrimination, which is difficult to realize in simulated data for small samples. The latent variable variance was fixed to one.

A total of 1,000 data sets were simulated under the 2PNO model (normal ogive metric) and under the 2PL model (logistic metric). Subsequently, item parameters were recovered by fitting the 2PL model using ML estimation (logistic metric) given the simulated 2PL item response data. The item parameters were also recovered by fitting the 2PNO model using MCMC estimation (normal ogive metric) given the simulated 2PNO item response data. This procedure was followed to avoid rescaling estimation results under the Probit model (normal ogive metric) to the Logit model (logistic metric) or the other way around—the metric used for data simulation was preserved in the parameter estimation. Therefore, the simulated levels of discrimination and difficulty were chosen to be the same for both models.

The 2PL and 2PNO models were identified by fixing the latent variable variance to one, in accordance to the simulation of the item response data. The average root-mean-squared error (RMSE) over 1,000 data sets was computed for the discrimination and difficulty parameters for both the 2PNO and the 2PL, which were directly comparable due to the design of the simulation study. Furthermore, for the 2PNO model, the average posterior standard deviation for the discrimination and difficulty parameters was computed and for the 2PL the average standard error. The results are given in Table 1, where the label SD is used to represent the average standard error (2PL) and the average posterior standard deviation (2PNO). For the 2PL, discrimination ( $a$ ) and difficulty ( $d$ ) estimates were given in the natural metric  $a_k\theta_i + d_k$ , and transformed to estimates in the metric  $a_k(\theta_i - b_k)$  with  $b_k = -d_k/a_k$ , which are referred to as IRT parameter estimates (R-package *mirt*). The difficulty ( $d$ ) and IRT difficulty estimates ( $b$ ) are given under the label 2PL in Table 1.

The IRT-BCSM was fitted on the simulated 2PNO data, since it is based on a Probit link function. The IRT-BCSM was identified by restricting the scale of the discrimination parameters. Therefore, the sum of squared item discriminations (squared Frobenius norm) was restricted to equal the simulated sum of squared item discriminations. This restriction also ensured that the item parameter estimates were on the same scale as the simulated item parameters. This restriction

TABLE 1.  
*Simulation Results for Item Parameters for IRT-BCSM, 2PNO (MCMC), and 2PL (ML) for Binary Item Response Data (Averaged Across 1,000 Replications for Different Sample Sizes)*

Statistic	IRT-BCSM		2PL (ML)			2PNO (MCMC)	
	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>	<i>d</i>	<i>a</i>	<i>b</i>
<i>n</i> = 1,000, <i>m</i> = 10							
RMSE	0.056	0.045	0.111	0.092	0.069	0.073	0.045
SD	0.057	0.052	0.113	0.098	0.073	0.074	0.052
<i>n</i> = 500, <i>m</i> = 10							
RMSE	0.067	0.064	0.163	0.140	0.099	0.113	0.067
SD	0.074	0.075	0.162	0.144	0.105	0.109	0.076
<i>n</i> = 100, <i>m</i> = 10							
RMSE	0.109	0.143	0.550	37.82	0.233	0.905	0.260
SD	0.134	0.142	0.433	>100	0.244	0.362	0.198
<i>n</i> = 50, <i>m</i> = 10							
RMSE	0.110	0.203	1.782	>100	0.529	3.621	1.101
SD	0.158	0.208	1.217	>100	0.498	1.004	0.513

*Note.* The reported RMSE and SD were both averaged across items. 2PNO (MCMC)=two-parameter normal ogive (Markov chain Monte Carlo); 2PL (ML)=two-parameter logistic regression (maximum likelihood); IRT-BCSM=IRT Bayesian covariance structure model.

was applied in each MCMC iteration by rescaling sampled item discriminations accordingly (Fox, 2010, pp. 88–89). The Gibbs sampler, described in Appendix C with the specified hyper prior values, was ran for 5,000 iterations and a burn-in period was used of 500 iterations. The shape and scale parameters of the inverse-gamma prior for  $\sigma_a^2$ ,  $\sigma_b^2$ , and  $\sigma_k^2$  were set at  $g_1 = 0.01$  and  $g_2 = 0.01$ , respectively.

None of the data sets were discarded in the analysis. Furthermore, all estimation results were used in the calculation of the statistics (RMSE and SD). Therefore, estimation results were included of MCMC runs for the 2PNO, which showed poor convergence and/or low effective sample sizes. For small samples, in some occasions MML estimation produced IRT estimates which showed a large deviation with the true values. The IRT-BCSM showed good convergence and sufficiently high effective sample sizes for all conditions and replicated data sets. By including all estimation results, the estimation problems for small samples is reflected in the computed statistics.

The results in Table 1 show that the (average) RMSE for the discrimination and difficulty parameters is smallest for the IRT-BCSM, when the sample size equals 1,000 and the prior influence is minimal. For the difficulty parameters, the RMSE of the IRT-BCSM is comparable to that of the 2PNO (MCMC). It can be

seen that the SEs for the 2PL estimates are higher, which automatically leads to a higher RMSE. When the sample size is decreased to  $n = 500$ , the IRT-BCSM still has the lowest RMSE, and the increase in RMSE is (substantially) higher for the 2PL and 2PNO models. For  $n = 100$ , the RMSEs of the discrimination and difficulty parameters of the 2PL and the 2PNO show problematic estimation results. The difficulty parameter estimates ( $b$ ) for the 2PL were unstable, which led to a rapid increase of the RMSE and the SD. For the 2PL, estimated discriminations close to zero led to substantial bias in the  $b$ -estimates. The  $d$ -estimates were more stable and produced much more acceptable results than the  $b$ -estimates. The difficulty estimates ( $d$ ) (in the natural metric) were slightly better than those of the 2PNO. However, the RMSE and SD of the discrimination and difficulty parameter estimates for the 2PL and 2PNO are higher and no longer in line with those of the IRT-BCSM.

For the IRT-BCSM, the estimates remain good and show an acceptable increase in RMSE and  $SD$ . When reducing the sample size from 500 to 100, the increase in RMSE of the IRT-BCSM is partly reduced by shrinkage effects implied by prior distributions for the item parameters. When the sample size is reduced, the discrimination and difficulty estimates each become more alike due to their priors leading to more stable estimation results, which follows from the borrowing-strength principle. This shrinkage effect is not present in the 2PL and the 2PNO estimates, and this partly explains the rapid increase in RMSEs when reducing the sample size. When the sample size is reduced to 50, the RMSE and SD remained stable for the IRT-BCSM. For the other models, the RMSEs and SDs show that the estimation results are inaccurate with (very) low precision.

For the IRT-BCSM, when the sample size decreases to  $n = 50$ , the RMSE and  $SD$  of the discrimination parameters become smaller than those of the difficulty parameters. One reason for this is that for really small sample sizes, there is hardly any information to distinguish discriminations between items, and the estimated discriminations become more similar. The discriminations were sampled from a normal distribution with standard deviation 0.10 and mean 0.80. Thus, the estimation results for RMSE and  $SD$  show a convergence toward these (default) values. The standard deviation to generate difficulty parameters was set at 0.50, and this leads to a higher RMSE and  $SD$  than those of the discriminations, when the difficulty estimates are shrunk more toward a common level.

A second reason is that the discrimination parameter is no longer a component of the linear term in the IRT-BCSM, but the difficulty parameter still is. Therefore, the difficulty estimates behave comparable to other programs and only doing better for smaller sample sizes. The discrimination parameter is a covariance component, and the estimation information does not depend on the recovery of the latent variable. For instance, estimation results are less sensitive to the shape of the frequency distribution of the response patterns, which can be typically problematic for small samples. For  $n = 1,000$ , it can be seen that the

TABLE 2.  
*Effective Sample Size Characteristics for the IRT-BCSM and the 2PNO (MCMC) for 1,000 Data Replications, With  $m = 10$  and Different Sample Sizes  $n$ , and 5,000 MCMC Iterations*

Parameter	IRT-BCSM			2PNO		
	$n = 50$	$n = 100$	$n = 500$	$n = 50$	$n = 100$	$n = 500$
Discrimination parameter						
Mean	1,462	1,372	1,072	167	353	452
Median	1,459	1,381	1,075	145	342	448
Min	405	462	546	1.13	2.12	172
Max	2,922	2,318	1,588	769	990	822
Skew	0.10	-0.10	-0.09	1.08	0.37	0.25
Kurtosis	-0.31	-0.17	0.18	1.46	-0.13	0.13
Difficulty parameter						
Mean	1,764	1,803	1,739	264	589	704
Median	1,795	1,832	1,746	171	550	692
Min	651	782	1,161	2.08	6.54	326
Max	2,684	2,732	2,293	2,226	2,005	1,244
Skew	-0.54	-0.38	-0.19	3.10	0.89	0.43
Kurtosis	1.35	0.25	1.16	13.04	1.48	0.19

*Note.* Reported statistics were averaged across items. 2PNO (MCMC)=two-parameter normal ogive (Markov chain Monte Carlo); 2PL (ML)=two-parameter logistic regression (maximum likelihood); IRT-BCSM=IRT Bayesian covariance structure model.

discrimination estimates are better than those of mirt (ML) and sirt (Gibbs sampling), but estimation results of the difficulty estimates are more similar.

*4.1.1 Study 1: MCMC Performance.* In Table 2, the characteristics of the distribution of the effective sample sizes (for 5,000 MCMC iterations) for the discrimination parameters and the difficulty parameters of the Gibbs sampler for the IRT-BCSM and for the 2PNO are given. It can be seen that for the IRT-BCSM, the reduction in sample size from 500 to 50 did not have a serious impact on the distribution of effective sample sizes. For instance, the average and minimum effective sample size remained sufficiently high in all sample size conditions. This means that the Gibbs sampler for the IRT-BCSM generated an informative sample for estimation for each item parameter for all sample sizes. This was not the case for the Gibbs sampler for the 2PNO. For the discrimination parameter, the minimum effective sample decreased from 172 ( $n=500$ ) to 1.13 ( $n=50$ ), and the median from 448 to 145. To reach a minimum effective sample size of 400—then the posterior standard error of the mean is less than 5% of the posterior standard deviation of the parameter—at least a total of 2.3 to 354 times 5,000 (11.5K

to 1.770K) MCMC iterations should have been made. When reducing the sample size, the efficiency of Gibbs sampler for the 2PNO was decreasing rapidly. For the 2PNO, the effective sample size distribution for the difficulty parameters is positively skewed, and the level of positive skewness increased when reducing the sample size. Thus, for a sample size of 50, the median effective sample size is 170, but the positive skewness reveals that there are MCMC samples with acceptable to high effective sample sizes. For the IRT-BCSM, this appearance is reversed. The negative skew reveals that there are some (relatively) low effective sample sizes but that most of the MCMC samples show a high effective sample size. It can be concluded that the Gibbs sampler for the IRT-BCSM is very efficient and also operates well for smaller sample sizes. The Gibbs sampler for the 2PNO is (very) inefficient, and for small samples, million(s) of MCMC iterations are required to obtain an acceptable effective sample size of 400.

#### 4.2 Study 2: Parameter Recovery

In this study, a total of 1,000 data sets were generated under the IRT-BCSM for sample sizes of  $n=50$ , 100, and 500 and for  $m=5$ , and 10. For each data set, item discrimination parameters were sampled from a normal distribution with mean 0.80 and standard deviation 0.25, and difficulty parameters were sampled from a standard normal distribution. The common covariance  $\tau$  was varied from zero (no person factor support by the data) to one (data support with random factor variance equal to one). The Gibbs sampler (Appendix C) was used to estimate the model parameters using 5,000 iterations for estimation, which led to an effective sample size of at least 400. The IRT-BCSM was identified by restricting the sum of squared sampled discriminations to equal the sum of squared simulated discriminations in each MCMC iteration.

In Figure 1, the RMSE, based on 1,000 replications, averaged across 5 and 10 items for the discrimination, and difficulty parameters are plotted for different values of the covariance parameter  $\tau$  and for  $n=50$  and  $n=100$ . For a substantial common covariance,  $\tau \geq .80$ , there was a reduction in the RMSE for the discrimination parameters, when increasing the number of items from 5 to 10 and when increasing the sample size from  $n=50$  to  $n=100$ . When the common covariance was small to zero,  $\tau \leq .10$ , increasing the number of items or the sample size from  $n=50$  to  $n=100$ , did not lead to a reduction in RMSE of the discrimination parameters. For the difficulty parameters, the RMSE decreased for all values for  $\tau$ , when increasing the number of items and when increasing the sample size from  $n=50$  to  $n=100$ . When increasing the sample size, the reduction in RMSE was higher when  $\tau$  was lower. The item response data became more informative about the difficulty parameters when the data is less correlated. The data became more informative about the discrimination parameters when the data was more

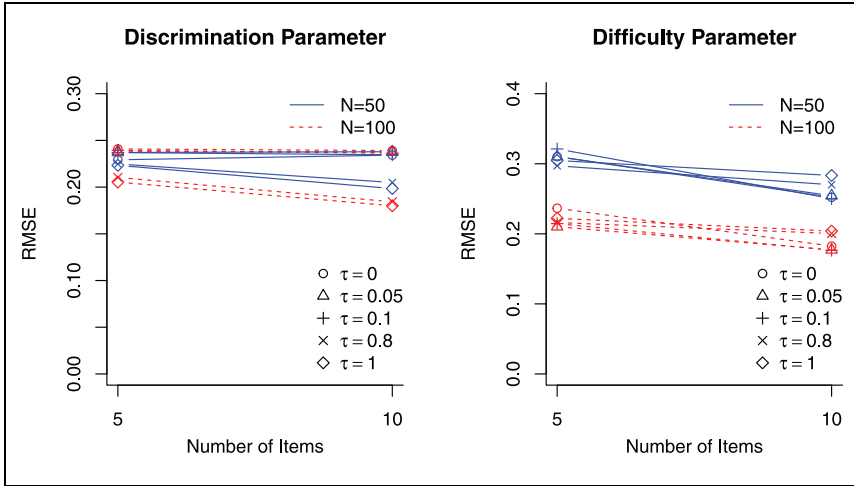


FIGURE 1. Average RMSE for discrimination and difficulty parameters across sample sizes and  $\tau$  values (averaged across 1,000 (binary) data replications).

correlated. The RMSEs show that both item parameters were accurately estimated for small sample sizes.

For small sample sizes, the RMSE for the discriminations went up to around 0.25, which resembles the standard deviation with which the discrimination parameters were generated for simulating data. Again, for small samples and for  $\tau$  (nearly) zero, the estimated discriminations are more alike and the RMSE converges to the standard deviation in generated discrimination parameters. The RMSEs represented in Figure 1 are also higher than those in Table 1 for similar sample sizes, since the standard deviation of generated discriminations in study 1 was 0.10, where it was 0.25 in study 2. The standard deviation of generated difficulty parameters was 1 in study 2, and 0.50 in study 1, which also led to higher RMSEs in study 2 for the estimated difficulty parameters.

For  $\tau$  (nearly) zero, the item discriminations can still be estimated, since the posterior distribution of  $\tau$  has its mode around zero, but also supports values around zero. This makes it still possible to sample discrimination values. The sampled discriminations are much alike, since there is hardly any between-item information. The identification rule restricts the discriminations to an average level, which resembles the simulated average level. The results for  $\tau$  (nearly) zero show that the IRT-BCSM can always be fitted, and return interpretable results, even when the response data does not support a (person) factor.

TABLE 3.  
*Simulation Results for  $\tau$  for IRT-BCSM for Continuous and for Binary Item Response Data (Averaged Across 1,000 Replications for Different Sample Sizes)*

True $\tau$	Estimate $\tau$ (SD)			
	$n = 100$		$n = 50$	
	$m = 10$	$m = 5$	$m = 10$	$m = 5$
Continuous				
0.00	0.00 (0.02)	0.01 (0.05)	0.01 (0.04)	0.02 (0.08)
0.05	0.05 (0.03)	0.05 (0.06)	0.06 (0.05)	0.07 (0.09)
0.10	0.10 (0.04)	0.10 (0.07)	0.11 (0.06)	0.12 (0.10)
0.80	0.82 (0.14)	0.81 (0.16)	0.82 (0.20)	0.83 (0.24)
1.00	1.00 (0.17)	1.01 (0.19)	1.01 (0.24)	1.02 (0.28)
Binary				
0.00	0.00 (0.05)	0.00 (0.14)	0.01 (0.07)	0.00 (0.22)
0.05	0.06 (0.06)	0.04 (0.15)	0.05 (0.09)	0.03 (0.24)
0.10	0.10 (0.07)	0.09 (0.18)	0.10 (0.10)	0.08 (0.28)
0.80	0.81 (0.22)	0.84 (0.44)	0.80 (0.31)	0.78 (0.71)
1.00	1.00 (0.27)	1.08 (0.55)	1.00 (0.38)	0.97 (0.84)

Note. Posterior mean ( $m = 10$ ) and median ( $m = 5$ ) estimates of  $\tau$  with posterior SD in brackets. IRT-BCSM = IRT Bayesian covariance structure model; SD = standard deviation.

In Table 3, the estimates of the common covariance parameter  $\tau$ , averaged across 1,000 data replications, are given. For  $m = 10$ , the posterior mean is used as an estimator, and for  $m = 5$ , the posterior median. The posterior distribution of  $\tau$  is more positively skewed when reducing the number of items. Then, the posterior mean tends to overestimate more easily the true value of  $\tau$  than the posterior median. For continuous outcome data (see Appendix B), the posterior estimates are close to the true values for all sample sizes. The posterior standard deviation increases when the true value of  $\tau$  increases, when the number of items ( $m$ ) decreases, and when the number of persons ( $n$ ) decreases. For binary data, the posterior estimates are also close to the true values even for the smallest sample size, but as expected the posterior standard deviations are higher than those for the continuous data. For the smallest sample size, the posterior standard deviations are increasing rapidly for increasing  $\tau$ . For other sample sizes, the posterior standard deviations are still quite acceptable.

## 5. Discussion

For small samples, a latent-variable-free IRT model can improve the accuracy of the estimation results. The IRT-BCSM represents a difficulty parameter in the mean part of the model, and a structured covariance matrix represents the

common dependence among clustered item responses, in which discrimination parameters modify the item-specific covariances. A Gibbs sampling algorithm is proposed, which results in efficient sampling of the model parameters. The IRT-BCSM does not include a random person factor to improve model fitting for small samples. In the IRT model, the person factor levels are assumed to be normally distributed and the level of deviation from normality is associated with the level of explained variance by the factor. The more the distribution of (realized) person factor levels deviate from a normal distribution, the poorer the fit of an IRT model, and less variance is explained by the random person factor. For small samples, there is obviously a higher risk that the distribution of the random factor person deviates from a normal distribution than for large samples. In the IRT-BCSM, the discrimination parameters are covariance parameters and do not operate as a slope parameter for the latent variable, which makes the estimation of discrimination parameters easier.

IRT defines a relationship between a person's overall performance and their individual performances on the measure's items. For the IRT-BCSM, the target is the relationship between the individual performances on the items. Although a construct measurement can still be estimated, this is a byproduct, since it is not an intrinsic component of the model and not the target object of estimation. However, the IRT-BCSM can improve the estimation of construct(s) for small samples and reduce the required test length and the time to complete a study. Further research is required to examine the performance of the IRT-BCSM in these areas. Other extensions in the direction of test equating, modeling construct multidimensionality, and model fit assessment require more research. An interesting component is the functioning of the discrimination parameters in the IRT-BCSM, since they determine the item-specific relationship of each item in connection to the others. This is relevant to determine the reliability of a scale to assess the contribution of each item and the construction of a scale. The leading principle is the eigenvalue of the outer product of the discrimination parameters, which is a key component in the estimation of the IRT-BCSM.

### *5.1 Generalizations of IRT-BCSM*

The IRT-BCSM represents a two-parameter IRT model, but other IRT models can be constructed in a similar way. The one-parameter IRT model can be represented as a cross-classified covariance structure model, with only two covariance parameters. Assuming a normal prior distribution for the item difficulty parameters with mean  $\mu_b$  and variance  $\sigma_b^2$ , the difficulty parameters can be integrated out. This leads to a multivariate normal distribution for the latent responses with a general mean  $\mu_b$  and a structured covariance matrix:

$$\Sigma = \tau(\mathbf{I}_n \otimes \mathbf{J}_m) + \sigma_b^2(\mathbf{J}_n \otimes \mathbf{I}_m) + (\mathbf{I}_n \otimes \mathbf{I}_m),$$

where  $\otimes$  is the Kronecker product and  $\mathbf{J}_m$  is a matrix of ones of dimension  $m$ . The estimation of  $\tau$  and  $\sigma_b$  provides the person and item factor variance, respectively. The number of parameters is not increasing when including data from new items or new persons, which makes it suitable for small data.

### 5.2 Multilevel IRT-BCSM

For a multilevel IRT (MLIRT) model (Fox, 2010), persons are assumed to be clustered by a (higher) factor variable. The MLIRT model with random factors for  $g$  groups (index  $j$ ) and for  $n$  persons (index  $i$ ) in each group has a mean structure of  $a_{k1}\theta_{ij} + a_{k2}\theta_j - b_k$ . When integrating out both random factors, the structured covariance matrix for the latent response variables is given by

$$\Sigma = \tau_1(\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{a}_1\mathbf{a}_1') + \tau_2(\mathbf{I}_g \otimes \mathbf{J}_n \otimes \mathbf{a}_2\mathbf{a}_2') + (\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{I}_m), \quad (8)$$

where  $\tau_1$  and  $\tau_2$  represent the common covariance for the clustering by persons and by groups, respectively.

When extending the MLIRT model with random item difficulty parameters (Fox et al., 2020), with variance  $\sigma_{b_k}$ , the structured covariance matrix in Equation 8 is extended with an extra component,

$$\begin{aligned} \Sigma = & \tau_1(\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{a}_1\mathbf{a}_1') + \tau_2(\mathbf{I}_g \otimes \mathbf{J}_n \otimes \mathbf{a}_2\mathbf{a}_2') \\ & + (\mathbf{I}_g \otimes \mathbf{J}_n \otimes D(\sigma_b)) + (\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{I}_m), \end{aligned}$$

where  $D(\sigma_b)$  represents a diagonal matrix with elements  $\sigma_{b_k}$  ( $k = 1, \dots, m$ ).

The additive structure of the covariance matrix represents the clustering of item responses by the different factor variables. Therefore, posterior computation can be based on similar techniques as described in Section 3. The advantage is that complex IRT models can be fitted more efficiently due to a reduction in the number of model parameters—the incidental parameters are integrated out of the model.

Choosing an appropriate prior for the random factor variance has received much attention in the literature, since estimation results are sensitive to its choice (Gelman, 2006; van Erp & Browne, 2021). For instance, in the IRT model, the random factor variance determines the amount of pooling of information across observations and stimulates heterogeneity across persons for an increasing variance. The conjugate inverse-gamma prior has been a popular choice, but received criticism for being too informative in stimulating support of the random factor. In the IRT-BCSM, a shifted inverse-gamma prior is defined, which avoids over-estimation of the common covariance parameter, since it also allows for negative values—as long as the covariance matrix is positive definite—and can place sufficient (prior) density mass around zero. Furthermore, the shifted inverse-gamma prior is conjugate and enables efficient parameter sampling

through a Gibbs sampler, even when the true common covariance is zero. The prior for the discrimination parameters can be adjusted to reduce the pooling of information across items. This can be done by a uniform prior or by making the normal prior less informative by fixing the prior variance at a large value.

In summary, the efficient parameterization of the IRT-BCSM makes it particularly suitable for small samples. Furthermore, the shifted inverse-gamma prior also enables support for low to no random factor support by the data, which avoids numerical problems in fitting the model.

## Appendix A

### Posterior Computations

The posterior distributions of the model parameters, which have not been discussed in Section 3, are given. To ease the notation, a default inverse-gamma prior is used with shape and scale parameters,  $g_1$  and  $g_2$ , respectively, for different variance parameters:

- The discrimination parameters have a normal prior with mean  $\mu_a$  and variance  $\sigma_a^2$ . For the hyper priors, a normal inverse-gamma prior is used, with  $\sigma_a^2 \sim \mathcal{IG}(g_1, g_2)$  and  $\mu_a \sim \mathcal{N}(\mu_{a_0}, \sigma_a^2/n_{a_0})$ . The posterior distribution of  $\mu_a$  and variance  $\sigma_a^2$  is given by

$$\mu_a | \mathbf{a}, \sigma_a^2 \sim \mathcal{N}\left(\frac{n_{a_0}}{m + n_{a_0}} \mu_{a_0} + \frac{m}{m + n_{a_0}} \bar{a}, \frac{\sigma_a^2}{m + n_{a_0}}\right), \quad (9)$$

and

$$\sigma_a^2 | \mathbf{a} \sim \mathcal{IG}\left(g_1 + m/2, g_2 + \left(\sum_k (a_k - \bar{a})^2 + \frac{mn_{a_0}}{m + n_{a_0}} (\bar{a} - \mu_{a_0})^2\right) / 2\right), \quad (10)$$

respectively.

- The posterior distribution of the difficulty parameters follow from the conditional distribution of the data in Equation 3. The difficulty parameters have a normal prior with mean  $\mu_b$  and variance  $\sigma_b^2$ . The difficulty parameter  $b_k$  has a normal posterior distribution,

$$b_k | \mathbf{z}_k, \sigma_k^2, \mu_b, \sigma_b^2 \sim \mathcal{N}(\omega_b (\mu_b / \sigma_b^2 - n \bar{z}_k / \sigma_k^2), \omega_b), \quad (11)$$

with  $\omega_b = (n / \sigma_k^2 + 1 / \sigma_b^2)^{-1}$ .

- The item error variance,  $\sigma_k^2$ , has an inverse-gamma (hyper) prior with shape  $g_1$  and scale  $g_2$ , and an inverse-gamma posterior with shape  $g_1 + n/2$  and scale parameter  $g_2 + \sum_i (z_{ik} + b_k)^2/2$ .
- The prior parameters  $\mu_b$  and variance  $\sigma_b^2$  have a normal inverse-gamma hyper prior with  $\sigma_b^2 \sim \mathcal{IG}(g_1, g_2)$  and  $\mu_b \sim \mathcal{N}(\mu_{b_0}, \sigma_b^2/n_{b_0})$ . Then, the posterior distribution of  $\mu_b$  and variance  $\sigma_b^2$  is given by

$$\mu_b | \mathbf{b}, \sigma_b^2 \sim \mathcal{N}\left(\frac{n_{b_0}}{m + n_{b_0}} \mu_{b_0} + \frac{m}{m + n_{b_0}} \bar{b}, \frac{\sigma_b^2}{m + n_{b_0}}\right), \quad (12)$$

and

$$\sigma_b^2 | \mathbf{b} \sim \mathcal{IG}\left(g_1 + m/2, g_2 + \left(\sum_k (b_k - \bar{b})^2 + \frac{mn_{b_0}}{m + n_{b_0}} (\bar{b} - \mu_{b_0})^2\right) / 2\right), \quad (13)$$

respectively.

## Appendix B

### Continuous Data

For continuous item response data, the item response data are considered to be multivariate normally distributed with a structured covariance matrix. The item responses are clustered by respondents, which leads to a common covariance for item responses belonging to the same person. Furthermore, a common error variance parameter across items is included. It follows that

$$\begin{aligned} \mathbf{Z}_i &\sim \mathcal{N}(-\mathbf{b}, \Sigma_c) \\ \Sigma_c &= \tau \mathbf{a} \mathbf{a}^t + \sigma_e^2 \mathbf{I}_m. \end{aligned} \quad (14)$$

The inverse of the covariance matrix  $\Sigma_c$  (Sherman-Morrison formula) is given by

$$\Sigma_c^{-1} = \sigma_e^{-2} (\mathbf{I}_m - \sigma_e^{-2} \mathbf{a} \mathbf{a}^t \tau / \lambda),$$

where  $\lambda = \sigma_e^{-2} \mathbf{a}^t \mathbf{a} + 1/\tau$ .

The mean and variance of the conditional normal distribution of  $Z_{ik}$  given  $\mathbf{Z}_{i,-k}$  follow from standard properties of the multivariate normal distribution and using the analytical form of the inverse of the covariance matrix. It follows that

$$z_{ik} | \mathbf{z}_{i,-k}, \mathbf{a}, \mathbf{b}, \tau \sim \mathcal{N}\left(-b_k + \frac{\sigma_e^{-2} a_k}{\lambda - k} \sum_{l \neq k} a_l (z_{il} + b_l), \sigma_z^2\right),$$

where  $\sigma_z^2 = \sigma_e^2 + a_k^2/\lambda_{-k}$  with  $\lambda_{-k} = \sigma_e^{-2} \mathbf{a}_{-k}^t \mathbf{a}_{-k} + 1/\tau$ . The distribution of item response vector  $k$ ,  $\mathbf{Z}_k$ , is also normal with mean  $-b_k$  and variance  $\sigma_k^2 = a_k^2 \tau + \sigma_e^2$ .

Given the conditional distribution for the response data, the posterior distributions of the item parameters and their prior parameters follow in a similar way as the ones derived for the binary item response data.

The posterior distribution of the common covariance  $\tau$  and error variance  $\sigma_e^2$  are derived from an eigenvalue decomposition of the covariance matrix  $\Sigma_c$ . Let  $\mathbf{D}$  represent the eigenvectors corresponding to the  $m$  eigenvalues. Then, the distribution of  $\mathbf{D}(\mathbf{Z}_i + \mathbf{b})$  is normal with mean zero and covariance matrix:

$$\begin{aligned} \mathbf{D}\Sigma_c\mathbf{D}' &= \mathbf{D}(\mathbf{a}\mathbf{a}'\tau + \sigma_e^2\mathbf{I}_m)\mathbf{D}' \\ &= \mathbf{a}'\mathbf{a}\tau\mathbf{K}_m + \sigma_e^2\mathbf{I}_m, \end{aligned}$$

with  $\mathbf{K}_m$  a single-entry precision matrix with a one at position (1, 1) and all other elements zero. Then,  $m - 1$  eigenvalues are  $\sigma_e^2$  and the other eigenvalue equals  $\tilde{\lambda} = \mathbf{a}'\mathbf{a}\tau + \sigma_e^2$ . A shifted inverse-gamma prior for  $\tau$  is defined, which includes a positive-definite restriction of the covariance matrix—restricting  $\tilde{\lambda}$  to be greater than zero:

$$p(\tau|\sigma_e^2, \mathbf{a}) \propto (\tau + \sigma_e^2/(\mathbf{a}'\mathbf{a}))^{-g_1-1} \exp\left(-\frac{g_2/(\mathbf{a}'\mathbf{a})}{\tau + \sigma_e^2/(\mathbf{a}'\mathbf{a})}\right),$$

where  $\sigma_e^2/(\mathbf{a}'\mathbf{a})$  is the shift parameter. Then, the posterior distribution for  $\tau$  is also a shifted inverse-gamma distribution:

$$p(\tau|\mathbf{z}, \mathbf{b}, \sigma_e^2, \mathbf{a}) \propto (\tau + \sigma_e^2/(\mathbf{a}'\mathbf{a}))^{-g_1-n/2-1} \exp\left(-\frac{\left(g_2 + \sum_i SS_i\right)/(\mathbf{a}'\mathbf{a})}{\tau + \sigma_e^2/(\mathbf{a}'\mathbf{a})}\right),$$

where  $SS_i = (\mathbf{D}'_i(\mathbf{z}_i + \mathbf{b}))^2$  and  $\mathbf{D}_i$  is the eigenvector corresponding to eigenvalue  $\tilde{\lambda}$ . It follows that the positive-definite restriction implies that  $\tau > -\sigma_e^2/(\mathbf{a}'\mathbf{a})$ .

Subsequently, the posterior distribution of  $\sigma_e^2$  is an inverse-gamma distribution with shape parameter  $g_1 + n(m - 1)$  and scale parameter  $g_2$  plus SSE =  $\sum_i \sum_{j \neq i} (\mathbf{D}'_j(\mathbf{z}_i + \mathbf{b}))^2$  using an inverse-gamma prior with shape  $g_1$  and scale  $g_2$ .

## Appendix C

### Gibbs Sampler for IRT-BCSM

The Gibbs sampler is defined for the IRT-BCSM for binary item response data.

The hyper prior parameters need to be set before running the algorithm. For the hyper prior of the discrimination parameters (Equations 9 and 10), the following values were used in the simulation study  $n_0 = 1$ ,  $b_0 = 1$ ,  $a_0 = 2$  for  $m \leq 5$ , and otherwise  $n_0 = 1$ ,  $b_0 = 0.01$ , and  $a_0 = 0.01$ . For the hyper prior of the difficulty parameters (Equations 12 and 13), in the simulation study, the values were set at  $n_0 = 1$ ,  $b_0 = 0.01$ , and  $a_0 = 0.01$ . The hyper prior parameters  $g_1$  and  $g_2$  were set to 0.01.

Initial values need to be set for discrimination parameters (all one in the simulation study), difficulty parameters (all zero in the simulation study), and common covariance parameter (zero in the simulation study). Initial latent responses can be sampled from a standard normal distribution. The prior parameters can be set to the following values  $\mu_a = 1$ ,  $\mu_b = 0$ ,  $\sigma_a^2 = .3$ , and  $\sigma_b^2 = 1$ , which was done in the simulation study.

Then, the Gibbs sampler consists of repeating the following steps.

1. Sample latent response data  $Z_{ik}$ , for each  $i$  and  $k$  from  $p(z_{ik}|y_{ik}, \mathbf{z}_{i,-k}, \mathbf{a}, \mathbf{b}, \tau)$  in Equation 4 and truncated to the domain  $z_{ik} \leq 0$  if  $y_{ik} = 0$  or  $z_{ik} > 0$  if  $y_{ik} = 1$ .
2. Sample item difficulties for each item  $k$  from  $p(b_k|\mathbf{z}_k, \sigma_k^2, \mu_b, \sigma_b^2)$  in Equation 11.
3. Sample difficulty population mean and variance from  $p(\mu_b|\mathbf{b}, \sigma_b^2)$  in Equation 12 and  $p(\sigma_b^2|\mathbf{b})$  in Equation 13, respectively.
4. Sample covariance parameter  $\tau$  from  $p(\tau|\mathbf{z}, \mathbf{b}, \mathbf{a})$  in Equation 7.
5. Sample item error variance  $\sigma_k^2$  for each  $k$  from an inverse-gamma distribution with shape  $g_1 + n/2$  and scale parameter  $g_2 + \sum_i (z_{ik} + b_k)^2/2$ .
6. Sample item discriminations  $a_k$  for each  $k$  from  $p(a_k|\mathbf{z}_k, \mathbf{x}_k, b_k, \mu_a, \sigma_a^2, \sigma_a^2)$  in Equation 5, which can be truncated from below to avoid sampling non-positive discriminations. Rescale sampled discrimination values such that the sum of squared item discriminations (squared Frobenius norm) equals a constant (identification rule).
7. Sample discrimination population mean and variance from  $p(\mu_a|\mathbf{a}, \sigma_a^2)$  in Equation 9 and  $p(\sigma_a^2|\mathbf{a})$  in Equation 10, respectively.

### *Shifted Inverse-Gamma Distribution*

Fox et al. (2017) introduced the shifted inverse-gamma distribution without any non-negativity restrictions, with the density function given by

$$\text{shifted-IG}(x; g_1, g_2, \sigma) = \frac{g_2^{g_1}}{\Gamma(g_1)} (x + \sigma)^{-(g_1 + 1)} \exp(-g_2/(x + \sigma)),$$

where  $\Gamma$  is the gamma function,  $g_1 > 0$  is the shape parameter and  $g_2 > 0$  the scale parameter. The parameter  $\sigma \in \mathbb{R}$  is called the shift parameter and the distribution

has support  $(-\sigma, \infty)$ . When  $\sigma = 0$ , the density equals an inverse-gamma density, denoted by  $\mathcal{IG}(x; g_1, g_2)$ . The corresponding distributions are denoted shifted  $-\mathcal{IG}(g_1, g_2, \sigma)$  and  $\mathcal{IG}(g_1, g_2)$ , respectively.


### Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author received no financial support for the research, authorship, and/or publication of this article.

### ORCID iD

Jean-Paul Fox  <https://orcid.org/0000-0002-0058-1496>

### References

- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*(3), 251–269. <https://doi.org/10.2307/1165149>
- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. <https://doi.org/10.1080/01621459.1993.10476321>
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. <https://doi.org/10.1007/BF02293801>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Sage.
- De Ayala, R. J. (2009). *Theory and practice of item response theory*. Guilford Publications.
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer. <https://doi.org/10.1007/978-1-4419-0742-4>
- Fox, J.-P., Koops, J., Feskens, R., & Beinhauer, L. (2020). Bayesian covariance structure modelling for measurement invariance testing. *Behaviormetrika*, *47*, 385–410. <https://doi.org/10.1007/s41237-020-00119-3>
- Fox, J.-P., Mulder, J., & Sinharay, S. (2017). Bayes factor covariance testing in item response models. *Psychometrika*, *82*(4), 979–1006. <https://doi.org/10.1007/s11336-017-9577-6>
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–534. <https://doi.org/10.1214/06-BA117A>
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

- Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior distribution variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279–291.
- Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. Springer.
- Kalbfleisch, J. D., & Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society. Series B (Methodological), 32*(2), 175–208. <http://www.jstor.org/stable/2984524>
- König, C., Spoden, C., & Frey, A. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement, 44*(4), 311–326. <https://doi.org/10.1177/0146621619893786>
- Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics, 95*, 391–413. [https://doi.org/10.1016/S0304-4076\(99\)00044-5](https://doi.org/10.1016/S0304-4076(99)00044-5)
- Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 44*(2), 226–233. <https://doi.org/10.1111/j.2517-6161.1982.tb01203.x>
- Reckase, M. D. (2009). Estimation of item and person parameters. In M. D. Reckase (Ed.), *Multidimensional item response theory* (pp. 137–178). *Statistics for Social and Behavioral Sciences*. Springer. [https://doi.org/10.1007/978-0-387-89976-3\\_6](https://doi.org/10.1007/978-0-387-89976-3_6)
- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133–144. <https://doi.org/10.1111/j.1745-3984.1990.tb00738.x>
- Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of Multilog. *Applied Psychological Measurement, 16*(1), 1–16. <https://doi.org/10.1177/014662169201600101>
- van Erp, S., & Browne, W. J. (2021). Bayesian multilevel structural equation modeling: An investigation into robust prior distributions for the doubly latent categorical model. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(6), 875–893. <https://doi.org/10.1080/10705511.2021.1915146>

### Author

JEAN-PAUL FOX is an associate professor at the Faculty of Behavioural, Management and Social Sciences, University of Twente, PO Box 217, 7500 AE Enschede, the Netherlands; e-mail: [j.p.fox@utwente.nl](mailto:j.p.fox@utwente.nl). His research interest is in Bayesian response modeling and Bayesian covariance structure modeling particularly in the context of large-scale surveys.

Manuscript received October 24, 2023

Revision received April 23, 2024

Accepted June 1, 2024