Article

Applied Psychological Measurement
2020, Vol. 44(4) 311–326
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621619893786
journals.sagepub.com/home/apm
SAGE

An Optimized Bayesian Hierarchical Two-Parameter Logistic Model for Small-Sample Item Calibration

Christoph König[1], Christian Spoden[2] and Andreas Frey[1,3]

Abstract

Accurate item calibration in models of item response theory (IRT) requires rather large samples. For instance, $N > 500$ respondents are typically recommended for the two-parameter logistic (2PL) model. Hence, this model is considered a large-scale application, and its use in small-sample contexts is limited. Hierarchical Bayesian approaches are frequently proposed to reduce the sample size requirements of the 2PL. This study compared the small-sample performance of an optimized Bayesian hierarchical 2PL (H2PL) model to its standard inverse Wishart specification, its nonhierarchical counterpart, and both unweighted and weighted least squares estimators (ULSMV and WLSMV) in terms of sampling efficiency and accuracy of estimation of the item parameters and their variance components. To alleviate shortcomings of hierarchical models, the optimized H2PL (a) was reparametrized to simplify the sampling process, (b) a strategy was used to separate item parameter covariances and their variance components, and (c) the variance components were given Cauchy and exponential hyperprior distributions. Results show that when combining these elements in the optimized H2PL, accurate item parameter estimates and trait scores are obtained even in sample sizes as small as $N = 100$. This indicates that the 2PL can also be applied to smaller sample sizes encountered in practice. The results of this study are discussed in the context of a recently proposed multiple imputation method to account for item calibration error in trait estimation.

Keywords

Bayesian, hierarchical models, item response theory, calibration, simulation, small samples

Item response theory (IRT) models such as the two-parameter logistic (2PL) model are currently the state of the art of measuring individual competences. Because of their complexity, however, they are associated with high sample size requirements. For instance, for accurate item calibration a minimum sample size of $N = 500$ is typically recommended for the 2PL (Baker, 1998; Liu & Yang, 2018). These sample size requirements pose a considerable challenge for applying the 2PL (or more complex models) to small-sample situations (De Ayala, 2009), such as

[1] Goethe University Frankfurt, Germany
[2] German Institute for Adult Education—Leibniz Centre for Lifelong Learning, Bonn, Germany
[3] Centre for Educational Measurement at the University of Oslo (CEMO), Norway

Corresponding Author:
Christoph König, Educational Psychology: Measurement, Evaluation and Counseling, Institute of Psychology, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 6, 60629 Frankfurt am Main, Germany.
Email: koenig@psych.uni-frankfurt.de

university exams or computerized adaptive tests, and items are calibrated with sample sizes smaller than recommended, introducing error in the subsequent estimation of trait scores (de la Torre & Hong, 2010; Feuerstahler, 2018).

To reduce item calibration error in small-sample IRT modeling, Bayesian approaches are proposed as alternatives to maximum likelihood (ML) estimation (Fox, 2010; Kim, 2001). The single-stage fully Bayesian estimation of IRT models, however, is criticized for being conceptually complex and computationally inefficient (Yang et al., 2012). Moreover, to increase the accuracy of item parameters in small samples, researchers are required to introduce prior information about the model parameters (or generally, about the population distribution of the parameters of interest) into the analysis (Swaminathan et al., 2003). When appropriate prior information is not available, a hierarchical approach to Bayesian estimation of IRT models offers a viable alternative; Swaminathan and Gifford (1985) and Mislevy (1986) were among the first to propose hierarchical versions of the 2PL (H2PL) model and to note their benefits for small-sample item calibration. Hierarchical Bayesian IRT models, such as the H2PL, exhibit a hierarchical structure of the prior distributions for the item parameters (Fox, 2010). The first level consists of a (usually multivariate) prior distribution for the vector of item parameters $\boldsymbol{\xi}$. The hyperparameters of this distribution, the vector of grand means of the item parameters $\boldsymbol{\mu}_{\boldsymbol{\xi}}$, and their variance–covariance matrix $\boldsymbol{\Sigma}$ (which contains the covariance of the item parameters and their variance components τ_α and τ_β), are not specified by the researcher directly but are given prior distributions themselves. These hyperprior distributions for $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ and $\boldsymbol{\Sigma}$ constitute the second level of the prior structure. This hierarchical structure yields more accurate parameter estimates in small samples than their nonhierarchical counterparts by pooling information across parameters of the same type, depending on τ_α and τ_β (Fox, 2010; Jackman, 2009). This beneficial characteristic was demonstrated for the H2PL, for instance, by Sheng (2013) and Natesan et al. (2016). Moreover, the hierarchical structure requires researchers to specify prior distributions only for the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ and $\boldsymbol{\Sigma}$. This is an important advantage because in nonhierarchical models, the benefits of the Bayesian approach in small samples can only be realized with adequate informative prior distributions (Sheng, 2010). Their specification, however, is not straightforward (Ames & Smith, 2018). Thus, utilizing a hierarchical approach alleviates this problem (Kim et al., 1994; Sheng, 2013). Nonetheless, the specification of prior distributions for $\boldsymbol{\Sigma}$ and τ_α and τ_β still requires careful consideration.

In the standard H2PL, $\boldsymbol{\Sigma}$ is commonly given a conjugate inverse Wishart prior distribution with $k \times k$ scale matrix S and degrees of freedom υ, where k equals the number of item parameters and $\upsilon > k - 1$. This is well known to be problematic (for a more detailed summary, see Alvarez et al., 2016) for three reasons: (a) uncertainty for all variances is controlled only by the hyperparameter υ; (b) if $v > 1$, the resulting marginal distribution for the variances has low density near zero, which biases associated estimates for variance components; and (c) the distribution contains a priori dependencies between correlations and variance components. The alternative is to separate covariance and variance components to give them individual prior distributions (Barnard et al., 2000). The use of the inverse gamma distribution as prior distribution for variance components, however, is discouraged in the recent Bayesian multilevel literature. Alternatives have been proposed in the form of the Cauchy and exponential distributions: both are heavy-tailed with higher mass around zero, compared to the inverse gamma distribution, which is known to be problematic when variance components are close to zero (Gelman, 2006; Polson & Scott, 2012). Using heavy-tailed distributions for variance components in hierarchical models in small-sample situations, however, has negative effects on the efficiency of the Markov chain Monte Carlo (MCMC) sampling (Betancourt & Girolami, 2013). Sampling inefficiencies may lead to bias in item parameter estimates, counteracting the reduction of item calibration error promised by the hierarchical approach. In the context of IRT models, these alternatives to the inverse gamma distribution became the focus of attention only recently (Liu & Yang, 2018; Sheng, 2017), while alternatives to the inverse Wishart distribution, or questions of sampling efficiency, were widely ignored.

The main assumption underlying this article is as follows. To utilize the full potential of the hierarchical approach for small-sample IRT modeling, an optimized H2PL is necessary which (a) increases the sampling efficiency when using heavy-tailed hyperprior distributions for τ_α and τ_β; (b) applies a separation strategy to $\boldsymbol{\Sigma}$ instead of the standard inverse Wishart distribution; and (c) avoids the inverse gamma distribution as hyperprior for τ_α and τ_β.

Thus, the goal of the following simulation study, and its primary contribution, is to investigate and quantify the combined effect of these optimizations on the accuracy of estimation of the variance components τ_α and τ_β, item parameters α_i and β_i, and trait scores θ_j in small-sample IRT modeling, compared to its standard inverse Wishart specification and its nonhierarchical counterpart. In addition, two limited-information estimators, namely, the unweighted and weighted least squares estimators (ULSMV and WLSMV), were included in the simulation as popular counterparts for latent variable modeling with categorical data. The results of the simulation study will provide answers to the question of whether the hierarchical approach to small-sample IRT modeling outlined above indeed offers an efficient way to estimate complex IRT models, yielding accurate parameter estimates even in smallest sample sizes. The optimized H2PL is described next.

The Optimized H2PL IRT Model

Let $y_{ij} \in \{0, 1\}$ be the response of person j to item i, θ_j be the ability of person j, and α_i and β_i be the discrimination and difficulty parameters of item i, respectively. The ability parameter is typically given a standard normal prior distribution, and the item parameters $\boldsymbol{\xi}_\mathbf{i} = \{\log \alpha_i, \beta_i\}$ have a joint multivariate normal prior with mean vector $\boldsymbol{\mu}_{\boldsymbol{\xi}} = \{\mu_\alpha, \mu_\beta\}$ and variance–covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \tau_\alpha & \sigma_{\beta\alpha} \\ \sigma_{\alpha\beta} & \tau_\beta \end{pmatrix}$, where τ_α and τ_β are the variance components, and $\sigma_{\alpha\beta}$ and $\sigma_{\beta\alpha}$ are the covariances of the item parameters. The log-transformation of α_i makes it possible to sample the transformed discrimination and difficulty parameters as correlated draws from a bivariate normal distribution (Glas & van der Linden, 2003). If the logit of a function x is defined by

$$\text{logit} = \frac{\exp(x)}{1 + \exp(x)}, \tag{1}$$

then the first level of the optimized H2PL can formally be expressed as follows:

$$\Pr\left(y_{ij} = 1 | \theta_j, \alpha_i, \beta_i\right) = \text{logit}\left[\alpha_i\left(\theta_j - \beta_i\right)\right] \tag{2.1}$$

$$\theta_j \sim N(0, 1) \tag{2.2}$$

$$\widetilde{\boldsymbol{\xi}_\mathbf{i}} \sim N(0, 1), \tag{2.3}$$

where $\widetilde{\boldsymbol{\xi}_\mathbf{i}} \sim N(0, 1)$ is a vector of uncorrelated $z-$scores related to the item parameters.

Equation (2.3) implies a reparametrization of the H2PL to simplify the sampling process and to increase the efficiency of the MCMC sampler, which is commonly found to be restricted in models with highly correlated posterior distributions, such as hierarchical models, irrespective of the MCMC sampler used (Betancourt & Girolami, 2013; Papaspiliopoulos et al., 2007). Posterior distributions with correlated dimensions are frequently associated with convergence problems and low effective sample sizes (ESSs; Turner et al., 2013). The ESS indicates the number of independent samples from the typical set of the target distribution included in an MCMC chain (Annis et al., 2017). It is defined by $\mathrm{ESS} = (n/(1 + 2\sum_{l=1}^{\infty} \rho(l)))$, where n is the total number of samples in the chain, and $\rho(l)$ is the autocorrelation of two adjacent samples (Betancourt, 2018). Autocorrelation depends on the correlation in a joint posterior distribution and indicates sampling inefficiencies that negatively affect the ESS.

The noncentered parameterization of the optimized H2PL alleviates sampling inefficiencies in two steps following Betancourt and Girolami (2013) for general Bayesian hierarchical models. First, it removes the cross-level dependency of the vectors of correlated item parameters $\boldsymbol{\xi}_i$ and their grand means $\boldsymbol{\mu}_{\xi}$, which is present when $\boldsymbol{\xi}_i$ is sampled from a multivariate normal distribution $\boldsymbol{\xi}_i \sim \mathrm{MVN}(\boldsymbol{\mu}_{\xi}, \boldsymbol{\Sigma})$, by subtracting the grand means and factoring out the variance components τ_α and τ_β. Second, the reparameterization removes the remaining correlation between the item parameters $\boldsymbol{\xi}_i$ by utilizing the general fact that draws from a multivariate normal distribution can be obtained by a Cholesky decomposition of the correlation matrix $\mathbf{L}_{\boldsymbol{\Omega}}$ (with $\boldsymbol{\Omega} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is a lower triangular matrix). In the noncentered H2PL, for each item $i, i = 1, \ldots, I$, a vector of uncorrelated $z-$scores $\tilde{\boldsymbol{\xi}}_i = (\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_I)$ is drawn from a standard normal distribution. Each individual vector is then multiplied by $\boldsymbol{\Lambda}$, the diagonal matrix of variance components τ_α and τ_β, and the Cholesky factor $\mathbf{L}_{\boldsymbol{\Omega}}$ to obtain the vector of item parameters $\boldsymbol{\xi}_i$ for each item. The deterministic transformations $\boldsymbol{\xi}_i = (\boldsymbol{\Lambda}\mathbf{L}_{\boldsymbol{\Omega}}\tilde{\boldsymbol{\xi}}_i)^{\mathrm{T}}$, $\alpha_i = \mu_\alpha + \xi_{\alpha i}$, and $\beta_i = \mu_\beta + \xi_{\beta i}$ effectively remove all dependencies of the H2PL from the sampling process, leaving only the uncorrelated θ_j and $\tilde{\boldsymbol{\xi}}_i$ as actively sampled variables on the first level of the optimized H2PL. The resulting joint posterior distribution has a much more convenient form, which the MCMC sampler is able to explore more efficiently, yielding lower autocorrelations and a higher ESS, because the parameter space is uncorrelated. A Stan implementation of the optimized H2PL is provided in the supplementary material.

The second level of the optimized H2PL includes the hyperpriors for $\boldsymbol{\mu}_{\xi}$, that is, the grand means of the discrimination and difficulty parameters, the hyperprior for $\mathbf{L}_{\boldsymbol{\Omega}}$, and for the variance components τ_α and τ_β:

$$\mu_\alpha \sim N(0, 1) \tag{2.4}$$

$$\mu_\beta \sim N(0, 2) \tag{2.5}$$

$$\mathbf{L}_{\boldsymbol{\Omega}} \sim LKJ(4) \tag{2.6}$$

$$\tau_{\alpha,\beta} \sim \mathrm{Cauchy}(0, 1). \tag{2.7}$$

The separation strategy based on $\mathbf{L}_{\boldsymbol{\Omega}}$ in equations (2.6) and (2.7) follows Barnard et al. (2000) and is implemented to avoid the well-known problems of the inverse Wishart distribution as hyperprior for $\boldsymbol{\Sigma}$ (Alvarez et al., 2016). It eliminates the a priori dependencies between the variance components and the covariances and offers more flexibility in prior specification, that is, an increased control of the uncertainty associated with the variance components. In the optimized H2PL, $\mathbf{L}_{\boldsymbol{\Omega}}$ is given a $LKJ(\mathbf{L}_{\boldsymbol{\Omega}}|\eta)$ prior distribution with the shape parameter η (Lewandowski et al., 2009). For a $k \times k$ lower triangular Cholesky factor of a correlation matrix $\mathbf{L}_{\boldsymbol{\Omega}}$ and $\eta > 0$, this distribution is defined by $LKJ(\mathbf{L}_{\boldsymbol{\Omega}}|\eta) = \prod_{k=2}^{K} L_{kk}^{K-k+2\eta-2}$ (Stan Development Team, 2016). The shape parameter η controls the degree of information contained in the prior distribution; as $\eta \to \infty$, extreme correlations become less probable. This prior distribution is currently widely used in Bayesian analyses involving covariance matrices (a) because it provides direct control over how closely the sampled matrix resembles the identity matrix and (b)

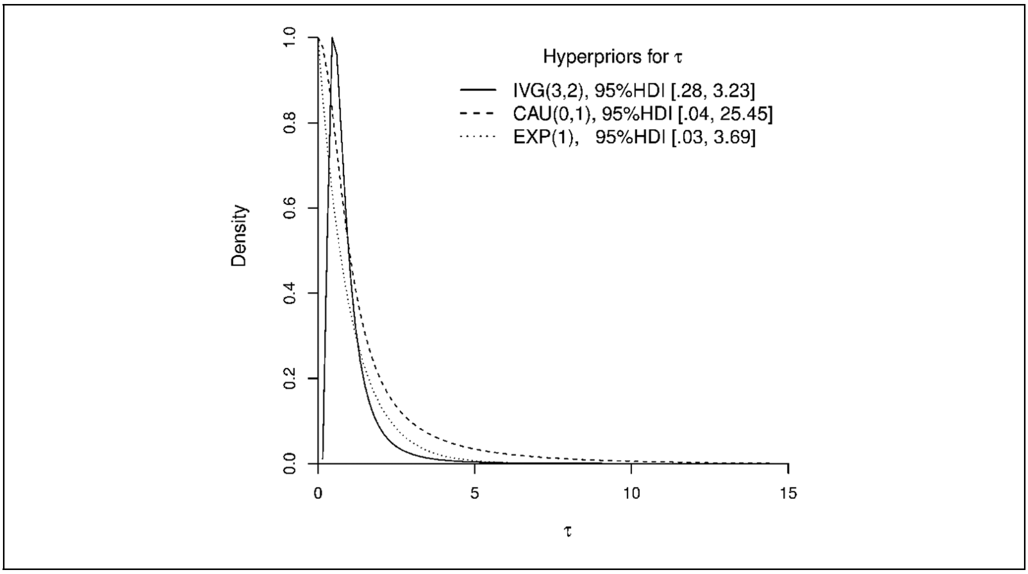

Figure 1. Densities of the IVG, CAU, and EXP distributions.
All three distributions are equivalently specified with μ = 1 and σ = 1. For each distribution, the 95% HDI is shown. Since the variance components τ_α and τ_β cannot be negative, but the Cauchy distribution has support on the real line, it is truncated at zero, that is, it is a half-Cauchy distribution.
Note. IVG = inverse gamma; CAU = Cauchy; EXP = exponential; HDI = highest density interval.

because of its numerical stability compared to the standard inverse Wishart distribution (Stan Development Team, 2016).

There are several alternatives regarding the choice of a weakly informative prior distribution for the variance components τ_α and τ_β. The inverse gamma distribution $\mathrm{IVG}(\tau_{\alpha,\beta}|a,b) = (b^a/\Gamma(a))\tau_{\alpha,\beta}{}^{-(a+1)}\exp(-(b/\tau_{\alpha,\beta}))$, with shape and scale hyperparameters $a, b>0$, is commonly used because of its conjugacy. However, if the variance component is estimated to be near zero, because of its relatively low mass around zero, inference is sensitive to the choice of the hyperparameters (Gelman, 2006). Thus, based on findings from the current methodological literature (Polson & Scott, 2012; Sheng, 2017), the optimized H2PL utilizes the Cauchy distribution $\mathrm{CAU}(\tau_{\alpha,\beta}|\mu,\sigma) = (1/\pi\sigma)(1/(1+((\tau_{\alpha,\beta}-\mu)/\sigma)^2))$, with location μ and scale σ. Due to its broad peak, it concentrates more mass around zero, leading to better performance around the origin and because of its thick tails, it also allows larger values if necessary (Polson & Scott, 2012). This might be problematic in nonlinear models with logit links, given possible floor and ceiling effects, because extreme values of the variance components are equally likely (McElreath, 2016). Based on the results of their simulation study on the utility of Cauchy prior distributions for logit link models, Ghosh et al. (2018) also state that for such (nonlinear) models, it may be necessary to consider alternatives to the heavy-tailed Cauchy distribution. The exponential distribution $\mathrm{EXP}(\tau_{\alpha,\beta}|b) = \beta\exp(-\beta\tau_{\alpha,\beta})$ with inverse scale $\beta>0$ is such a possible alternative. The peak around its mean is broader than that of the inverse gamma distribution, but thinner than that of the Cauchy distribution, and its tail is thinner, yielding estimates that are more conservative (McElreath, 2016). Figure 1 illustrates the difference in densities of these distributions, equivalently specified to match $\mu=1$ and $\sigma=1$. These weakly informative specifications can be found frequently in the context of the adaptive regularization of hierarchical models (McElreath, 2016).

Simulation Study

To examine the combined effect of the three optimizations, (a) sample size ($N = 50, 75, 100, 150, 200, 500$), (b) test length ($k = 25, 50$), and (c) specification (hierarchical and nonhierarchical) of the 2PL model were manipulated in a simulation study. The hyperprior distributions (inverse gamma, Cauchy, and exponential) and the parameterization (centered and noncentered) were nested in the specification factor. In total, the design consisted of $6 \times 2 \times 6 = 72$ cells. The design covered sample sizes typically regarded as suboptimal for item calibration under the 2PL because deriving accurate parameter estimates was shown to be problematic (De Ayala, 2009; Stone, 1992). The sample size of $N = 500$, which was considered the minimum sample size required for the 2PL, served as the baseline condition. Furthermore, the design covered test lengths that are commonly found in operational tests and prior research on Bayesian estimation of IRT models (Sheng, 2017). To give an even better indication of the performance of the optimized H2PL, it was furthermore compared to the standard inverse Wishart specification of the H2PL and to two popular limited-information estimators for categorical data (ULSMV and WLSMV).

Data Generation and Analysis

For each cell of the simulation design, 100 data sets were generated from a unidimensional 2PL model with correlated item parameters. Based on an analysis of descriptive statistics of item parameters from several large-scale assessments and based on recommendations from the literature, generating values for the variance components were set to $\tau_\alpha = 0.25$ and $\tau_\beta = 1$, and the correlation of the item parameters was set to $\rho_{\alpha,\beta} = .30$ (Fox, 2010). These generating values reflect variance components and dependencies of item parameters typically found in operational tests. Thus, item parameters were drawn from a multivariate distribution with mean vector $\boldsymbol{\mu}_\xi = \{0, 0\}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 0.0625 & 0.075 \\ 0.075 & 1.000 \end{pmatrix}$. This yielded typical item parameters (99% confidence intervals [CI] = $[0.47,\ 2.17]$ and $[-3.10,\ 3.08]$ of the generated discriminations and difficulties, respectively). Person parameters were drawn from a standard normal distribution $\theta_j \sim N(0, 1)$, generating a 99% CI = $[-3.11,\ 3.11]$ for the person parameters. Different sets of item and person parameters were drawn for each of the 100 data sets.

The centered H2PL was specified with $\boldsymbol{\xi}_\mathbf{i} \sim MVN(\boldsymbol{\mu}_\xi, \boldsymbol{\Omega})$ instead of equation (2.3). The equivalent specifications of the hyperprior distributions, as shown in Figure 1, represent weakly regularizing hyperprior distributions for variance components in general hierarchical models (McElreath, 2016). Given that $\tau_\alpha, \tau_\beta \geq 0$, the Cauchy distribution is a half-Cauchy distribution truncated at zero. The standard inverse Wishart H2PL was specified with $\theta_j \sim N(0, 1)$, $\boldsymbol{\xi}_\mathbf{i} \sim MVN(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma})$, $\mu_\alpha \sim N(0, 1)$, $\mu_\beta \sim N(0, 2)$, and $\boldsymbol{\Sigma} \sim IW(3, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. The nonhierarchical 2PL was specified with $\theta_j \sim N(0, 1)$,$\alpha_i \sim logN(0, 1)$, and $\beta_i \sim N(0, 2)$. These prior configurations are widely used in Bayesian IRT modeling (Fox, 2010; Levy & Mislevy, 2016).

Stan (Carpenter et al., 2017) and its R interface RStan (Stan Development Team, 2016) were used for Bayesian estimation. Four chains each with a length of 10,000 were set up with 5,000 burn-in cycles and a thinning interval of five, yielding a maximum ESS of 4,000 draws. Different random starting values were supplied to each of the four chains. Convergence was assessed using the Gelman–Rubin R-statistic (Gelman & Rubin, 1992), where $R<1.05$ indicated convergence. In the case of the centered specification of the H2PL, there was a small number of nonconvergent replications (under 10%). In the case of the noncentered specifications of the H2PL (and the standard inverse Wishart specification), all replications converged.

For the ULSMV and WLSMV estimation, *lavaan* (Rosseel, 2012) was used with "Theta" parameterization; since *lavaan* uses the probit link, loadings and thresholds were transformed into discriminations and difficulties using the correct formulas given by Paek et al. (2018). There were large numbers of nonadmissible replications (nonconvergent, negative variances, and not positive definite matrices) for both estimators across all sample sizes (up to 43%). Moreover, for $k = 50$, there were no admissible solutions for $N = 50$ and $N = 75$.

Dependent Measures

First, the sampling efficiency of the candidate hyperprior distributions for the variance components τ_α and τ_β was investigated to quantify the benefit of the noncentered parameterization of the optimized H2PL. Sampling efficiency was indicated by the average ESS of the variance components τ_α and τ_β and the average number of divergent transitions. Divergent transitions indicate that the MCMC chain was not able to adequately explore a region of high curvature in the posterior distribution (Betancourt, 2018). It was expected that the noncentered parameterization would increase the average ESS and eliminate divergent transitions; this pattern was expected to be more distinct for the Cauchy and exponential distributions, because of their thicker tails, compared to the inverse gamma distribution.

Second, the three hyperprior distributions of the optimized H2PL and the standard inverse Wishart specification of the H2PL were compared in terms of the accuracy of estimation of the variance components τ_α and τ_β. Accuracy of parameter estimation was indicated by the average bias (BIAS) and the root mean squared error (RMSE). Let τ be the true value of the variance component, and τ_r its estimate in the rth replication ($r = 1, \ldots, R$). Then $\text{BIAS}_\tau = \left(\sum_{r=1}^{R} (\tau_r - \tau)\right)/R$ and $\text{RMSE}_\tau = \sqrt{\left(\sum_{r=1}^{R} (\tau_r - \tau)^2\right)/R}$. Careful consideration must be given to the choice of hyperprior distribution because, given the borrowing principle (depending on τ_α and τ_β, information is pooled across parameters of the same type, yielding item parameter estimates balanced between their respective grand means and their item-specific estimates), bias in estimates of the variance components may lead to bias in item parameter estimates. It was expected that the inverse gamma distribution, due to its distinct peak, thin tail, and low mass in the region near zero, would perform worse than the Cauchy and Exponential distributions.

Third, the optimized H2PL was compared to the standard inverse Wishart specification, its nonhierarchical counterpart, and the ULSMV and WLSMV estimators in terms of the accuracy of estimation of the item parameters α_i and β_i and the accuracy of the trait scores θ_j estimated based on the estimated item parameters in the common two-stage approach. The BIAS and RMSE of α_i,β_i, and θ_j were averaged across items and persons, respectively, for each replication. To obtain the final BIAS and RMSE values, these replication-specific summary indices were averaged across replications. It was expected that the optimized H2PL would perform best. This implies that IRT models behave differently from general hierarchical models: typical values of α_i and β_i fall into a quite narrow range, which restricts their variances to be relatively small. Therefore, bias introduced by shrinkage might be negligible, and the increased amount of information available may fully contribute to an increase in the accuracy of estimation.

Results

Noncentering the H2PL Increases Sampling Efficiency

The noncentered parameterization is most beneficial for the optimized H2PL when its specification includes either the Cauchy or the exponential distribution as hyperprior for the variance components. As illustrated in Figure 2 (showing the average number of divergent transitions for $k = 25$), when using the inverse gamma distribution, the optimized H2PL exhibits hardly any divergent transitions, regardless of parameterization. Using either the Cauchy or the exponential distribution, the centered parameterization is associated with a considerable number of divergent transitions for all sample sizes of $N < 500$. When $k = 50$ (not shown), the average number of divergent transitions considerably increases for $N < 100$. Thus, the Cauchy and exponential distributions do not work well in smaller samples unless the H2PL is reparameterized. Noncentering the H2PL allows these alternative distributions to be utilized without restrictions in terms of validity of the parameter estimates when sample sizes are small.

The increase in sampling efficiency in terms of decreasing average numbers of divergent transitions is further reflected by the increase in the average ESS. There is an increase in the average ESS across all hyperprior distributions; it is most pronounced in the case of the Cauchy and exponential distributions, where the average ESS of the variance components is increased threefold for some sample sizes. Similar to the changes in the average number of divergent transitions, this indicates that the Cauchy and exponential distributions do not work well in the centered H2PL. Figure 2 illustrates the increase in average ESS for τ_α across parameterizations for all hyperprior distributions and $k = 25$; the increase is similar for $k = 50$. In the case of τ_β, the general pattern is similar, but the increase in the average ESS is not as large.

In sum, the Cauchy and exponential distributions do not work well in terms of sampling efficiency, compared to the inverse gamma distribution, unless the H2PL is reparameterized. Noncentering the optimized H2PL, however, effectively eliminates sources of bias in parameter estimates related to the efficiency of the sampling process. Thus, the following sections are based on results from the noncentered H2PL.

Using Alternatives to the Inverse Gamma Distribution Increases Accuracy of τ_α

Figure 3 illustrates differences in average BIAS and RMSE in estimates of the variance components between the candidate hyperprior distributions, compared to the standard inverse Wishart specification of the H2PL, across sample sizes and test lengths. Differences in average BIAS are most pronounced in the case of τ_α: except for $N = 500$ and $k = 50$, the inverse gamma distribution overestimates the variance of the item discriminations. The decreasing sample size introduces less bias in estimates of τ_α when using either the Cauchy or the exponential distribution. Overall, the optimized H2PL yields more accurate estimates of τ_α compared to the standard inverse Wishart specification of the H2PL across all test lengths and sample sizes. In the case of the average BIAS of τ_β, the candidate hyperprior distributions perform equally well.

Regarding τ_α, the advantages of the optimized H2PL over the standard inverse Wishart specification of the H2PL are also apparent in terms of RMSE. Differences between the inverse gamma, Cauchy, and exponential distributions emerge for sample sizes $N < 150$ for $k = 25$. The inverse gamma distribution exhibits a larger RMSE than the Cauchy or exponential distributions. For $k = 50$, the differences are negligible. In the case of τ_β, however, the inverse gamma distribution shows smaller RMSEs across sample sizes for $k = 25$. For $k = 50$, the largest differences in RMSE can be observed for sample sizes $N < 100$. The Cauchy distribution, however, shows the most consistent performance in terms of RMSE.

In sum, using either the Cauchy or the exponential distribution as hyperpriors for the variance components increases the accuracy of estimation for τ_α only. This leads, however, to a better adaptation of the item discrimination estimates to the amount of information in the data. Overall, the optimized H2PL outperforms the standard inverse Wishart specification of the H2PL in the case of τ_α across all test lengths and sample sizes.

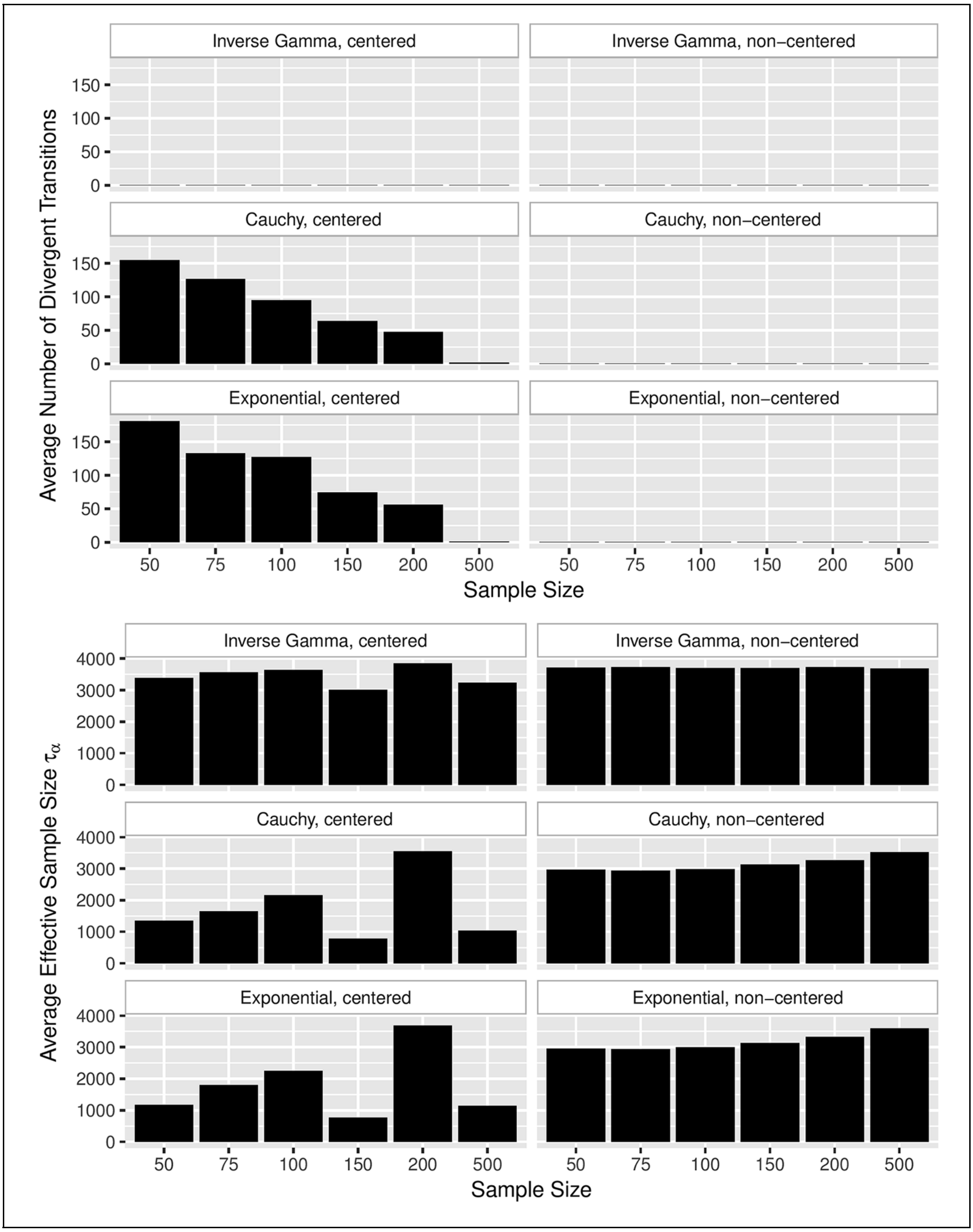

Figure 2. Sampling efficiency of the inverse gamma, Cauchy, and exponential distributions across parameterizations and sample sizes for k = 25.
Note. The nominal ESS of τ_α was 4,000.

The H2PL Yields Accurate Item Parameters and Trait Scores for Samples of N = 100

Figure 4 illustrates differences in average BIAS and average RMSE in item parameter estimates across sample sizes and test lengths between the optimized H2PL, its nonhierarchical

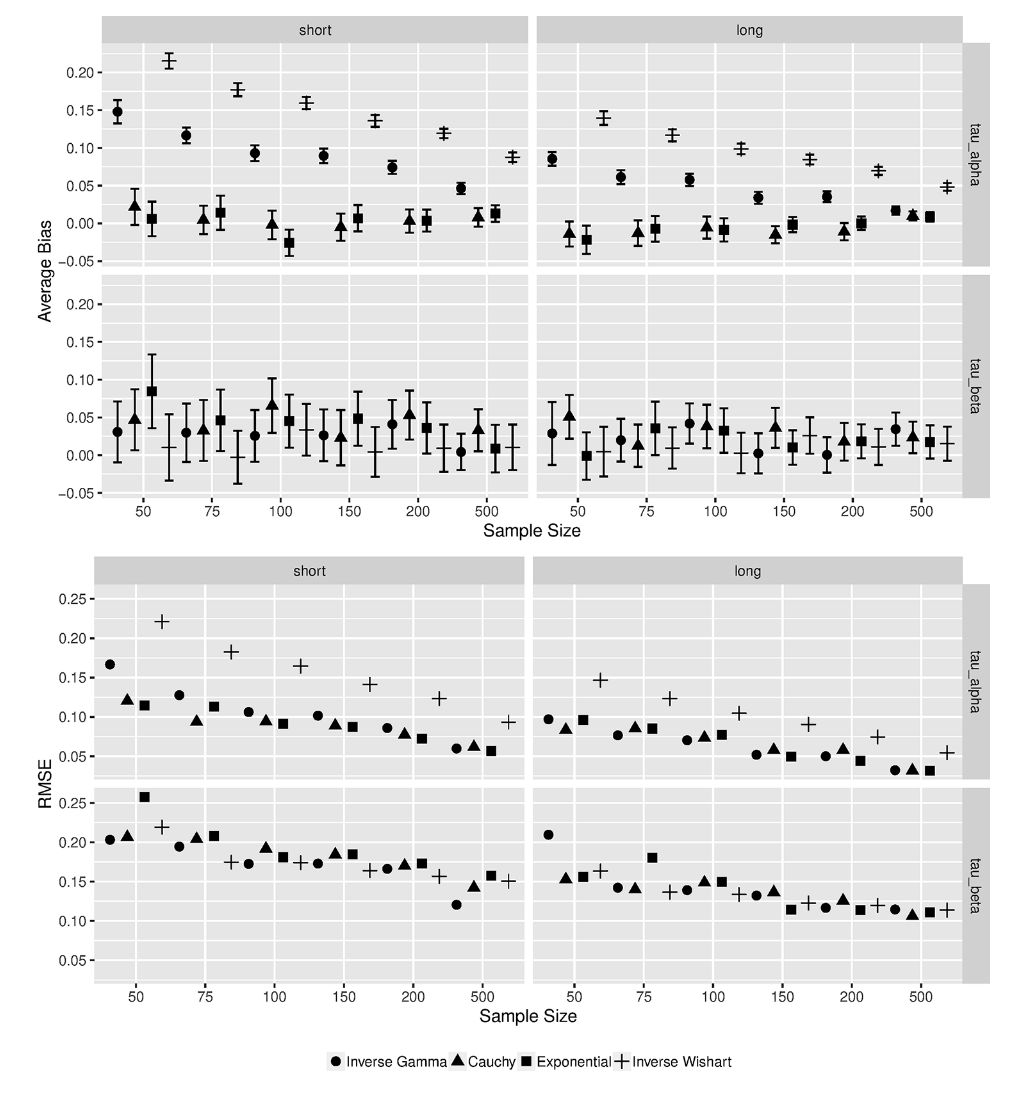

Figure 3. Differences in the accuracy of estimation of the variance components τ_α and τ_β between the inverse gamma, Cauchy, and exponential distributions across sample sizes for for k = 25 (short) and k = 50 (long).
Note. Error bars indicate ±2 SE.

counterpart, the standard inverse Wishart specification, and the ULSMV and WLSMV estimators. The nonhierarchical 2PL underestimates the item discrimination for all sample sizes and test lengths, except for $N = 50$ and $k = 25$. For the smallest sample sizes, there are also differences in average BIAS between the candidate hyperprior distributions in the optimized H2PL and its standard inverse Wishart specification. Both ULSMV and WLSMV estimators are outperformed by the Bayesian H2PL specifications when $N < 500$ for both test lengths. In the case of the item difficulty differences are less pronounced, both specifications perform equally well across sample sizes. Taking $N = 500$ as the nominal level, the average BIAS in item parameters

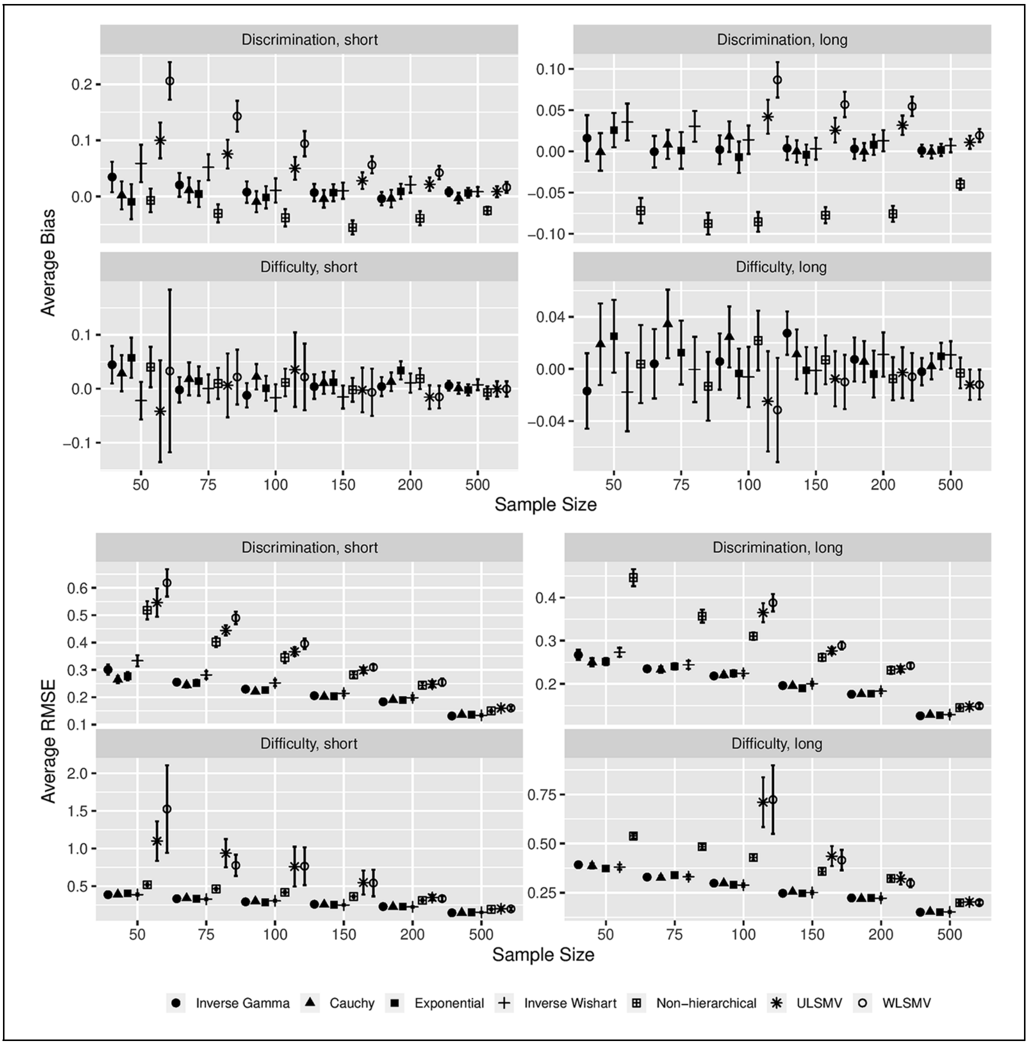

Figure 4. Differences in the accuracy of estimation of the item parameters between the optimized H2PL (with inverse gamma, Cauchy, and exponential distributions) and its standard inverse Wishart specification, its nonhierarchical counterpart, and the ULSMV and WLSMV estimators across sample sizes for k = 25 (short) and k = 50 (long).

Note. Error bars indicate ±2 SE.

does not considerably increase until $N = 100$ in the case of the optimized H2PL. In terms of average RMSE, the candidate hyperprior distributions perform equally well. Overall, differences in the average RMSE are most distinct between the hierarchical and nonhierarchical specifications (including the ULSMV and WLSMV estimators) for both item parameters across all sample sizes and test lengths: the hierarchical specifications consistently show smaller average RMSEs in item parameters.

Figure 5 illustrates if and how the increased accuracy of the item parameters translates into the accuracy of the trait scores for the Bayesian specifications of the 2PL. Overall, for both test lengths, the accuracy of the trait scores does not markedly decrease until $N = 100$, in terms of

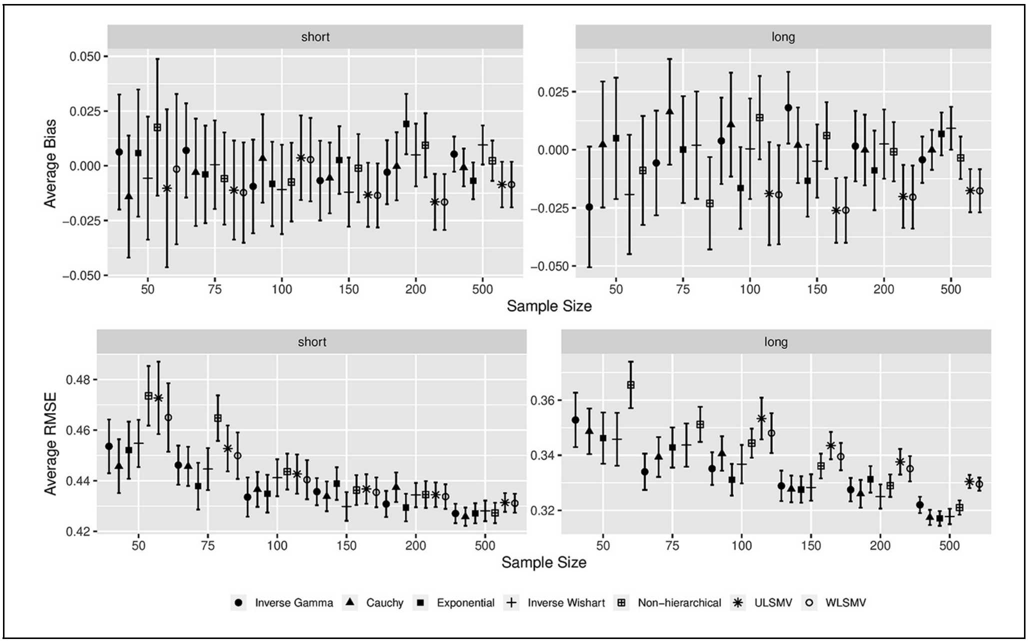

Figure 5. Differences in the accuracy of the trait scores between the optimized H2PL (with inverse gamma, Cauchy, and exponential distributions) and its standard inverse Wishart specification, its nonhierarchical counterpart, and the ULSMV and WLSMV estimators across sample sizes for *k* = 25 (short) and *k* = 50 (long).
Note. Error bars indicate ±2 SE.

average BIAS. There are no marked differences between the Bayesian specifications and the ULSMV and WLSMV estimators. Judging by the average RMSE, when $N<100,$ the accuracy of the trait scores becomes sensitive to the choice of specification; moreover, there is a slight increase in accuracy in the case of the optimized H2PL for $N<100$ and $k=25$, compared to its nonhierarchical counterpart. Compared to the ULSMV and WLSMV estimators, the average RMSE of the trait scores is lower in the case of the longer test length and $N>150$.

Discussion

The goal of this study was to investigate and quantify the effect of the optimized H2PL on the accuracy of estimation of the item parameters α_i and β_i and their variance components τ_α and τ_β in small-sample situations, and to investigate how this translates into the accuracy of trait scores θ_j. The optimized H2PL included (a) a noncentered parameterization, (b) the use of the Cholesky factor $\mathbf{L}_\Omega$ to separate variances and covariances, and (c) the use of the Cauchy and exponential distributions as alternative hyperprior distributions for the variance components. Noncentering the H2PL considerably increased the sampling efficiency in small sample sizes, especially when using the alternative hyperprior distributions for the variance components. It was further demonstrated that utilizing these alternative hyperprior distributions yields estimates of the variance components that are more accurate compared to the commonly used inverse gamma distribution. Moreover, when combining these elements in the optimized H2PL, this specification yields accurate item parameter estimates and trait scores even in sample sizes as small as $N=100$, which is considerably smaller than sample sizes recommended for item calibration or scoring (e.g., $N = 1,000$ or $N = 500$; Stone, 1992). As the 2PL is often regarded as a large-scale application, while typically only the simpler Rasch model is applied to sample sizes of approximately $N < 500$ (Stone & Yumoto, 2004), this finding is of practical importance since it shows that the 2PL can also be applied to sample sizes commonly encountered in practice.

This enhanced applicability of the 2PL can be attributed to the increased accuracy in the estimation of the item discrimination parameter and its associated variance component. The bias introduced by the underestimation of the item discrimination parameter in the standard, nonhierarchical 2PL across all sample sizes and test lengths has consequences for the estimation of trait scores. The accuracy of the trait score estimates includes, but is not limited to, item calibration error (Feuerstahler, 2018). The optimized H2PL reduces item calibration error in smaller sample sizes; as the item discrimination parameter is important for the calculation of the test information under the 2PL model, it is to be expected that the standard error of measurement of the trait scores is reduced as well. As the first indication of this effect, this study demonstrates the better performance of the optimized H2PL in terms of the average RMSE of the trait scores. It has to be noted that its performance is furthermore similar to both the ULSMV and the WLSMV estimators, where the trait scores are estimated without considering item calibration error.

Thus, the optimized H2PL may be most beneficial if applied to small-sample item calibration when item calibration error in the trait scores is to be accounted for. The common two-stage approach to trait estimation, where estimates of the item parameters are treated as true values without error, ignores the uncertainty carried over from the item calibration. Recently, a multiple-imputation-based approach has been proposed, in which m plausible item parameter values are drawn from a multivariate normal distribution with the ML-estimates of the item parameters as means and their asymptotic covariance matrix as scale (Yang et al., 2012). An alternative may be to draw m plausible item parameter values directly from their respective means and standard errors obtained under the optimized H2PL; the calculation of the asymptotic covariance matrix of the item parameters, based on the respective Fisher information matrix, would be no longer required (Liu & Yang, 2018). It may be promising to compare these two alternatives within the multiple-imputation-based approach to trait estimation, with a special focus on their performance in small samples. Nevertheless, the findings of this study indicate that the optimized H2PL could also be used in a single-stage approach to trait estimation; although item calibration error is taken into account, it yields an accuracy in the trait scores comparable to the ULSMV and WLSMV estimators. Its proposed use in the aforementioned two-stage approach, however, is conceptually easier to integrate into the standard operating procedures in applied testing situations (Yang et al., 2012).

The advantage of the optimized H2PL over its nonhierarchical counterpart in terms of bias in estimates of the item discrimination parameter is somewhat surprising. A potential explanation involves its variance component. Shrinkage of parameter estimates toward their grand means, hence their bias, depends on the variance of a given parameter. The increased accuracy of the item discrimination parameter might indicate that its variance is at a level where the bias, usually introduced by shrinkage, is outweighed by the increased amount of information available for the estimation of the item discrimination parameter. Thus, this result indeed points out the possibility that IRT models behave differently than general hierarchical models because typical values of α_i and β_i fall into a quite narrow range, which restricts their variances to be relatively small. Future simulations could address this general idea and remedy one limitation of this study: its focus on a single set of true values of the variance components. Although the choice of their generating values is based on operational item sets, it might be promising to investigate this pattern for different sets of generating values. Another limitation of this study is the focus on a single specification for the candidate hyperprior distributions. Although it was chosen to make them comparable and to take up recommendations from the current methodological literature, it may be fruitful to investigate how sensitive the results are to different specifications of the distributions, especially in small sample sizes. This may provide further evidence for their utility for small-sample IRT modeling.

Finally, the results of this study contribute to the growing body of literature discouraging the use of the inverse gamma distribution (Gelman, 2006; Polson & Scott, 2012). Even in a weakly informative specification, it overestimates the variance of the item discrimination parameter across almost all sample sizes and test lengths. The advantages of both the Cauchy and exponential distributions, as shown in this study, contribute to recent studies investigating these distributions as viable alternatives (Liu & Yang, 2018; Sheng, 2017). However, the use of either the Cauchy or the exponential distribution requires a reparameterization of the H2PL to ensure the validity of item parameter estimates. In summary, this study illustrates how to apply the 2PL model, usually considered a large-scale application, to small-sample situations.

Acknowledgments

The authors thank the Editor in chief Dr. John R. Donoghue and Dr. Hua-Hua Chang, the associate editors, and the anonymous reviewers for their valuable and helpful comments on their submitted manuscript.

Declaration of Conflicting Interests

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

Funding

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

ORCID iDs

Christoph König https://orcid.org/0000-0003-3172-7029
Christian Spoden https://orcid.org/0000-0002-2108-6152
Andreas Frey https://orcid.org/0000-0001-5334-9538

Supplemental Material

Supplementary material is available for this article online.

References

Alvarez, I., Niemi, J., & Simpson, M. (2016). Bayesian inference for a covariance matrix. *Annual Conference on Applied Statistics in Agriculture*, *26*, 71–82. https://arxiv.org/abs/1408.4050

Ames, A., & Smith, E. (2018). Subjective priors for item response models: Application of elicitation by design. *Journal of Educational Measurement*, *55*(3), 373–402. https://doi.org/10.1111/jedm.12184

Annis, J., Miller, B. J., & Palmeri, T. J. (2017). Bayesian inference with Stan: A tutorial on adding custom distributions. *Behavior Research Methods, 49*, 863–886. https://doi.org/10.3758/s13428-016-0746-9

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*(2), 153–169. https://doi.org/10.1177/01466216980222005

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*(4), 1281–1312.

Betancourt, M. (2018). *A conceptual introduction to Hamiltonian Monte Carlo*. https://arxiv.org/abs/1701.02434v2

Betancourt, M., & Girolami, M. (2013). *Hamiltonian Monte Carlo for hierarchical models*. https://arxiv.org/abs/1312.0906

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement*, *34*(4), 267–285. https://doi.org/10.1177/0146621608329501

Feuerstahler, L. M. (2018). Sources of error in IRT trait estimation. *Applied Psychological Measurement*, *42*(5), 359–337. https://doi.org/10.1177/0146621617733955

Fox, J.-P. (2010). *Bayesian item response modeling*. Springer.

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., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511. https://doi.org/10.1214/ss/1177011136

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*(4), 247–261. https://doi.org/10.1177/0146621603027004001

Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, *13*(2), 359–383. https://doi.org/10.1214/17-BA1051

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement*, *25*(2), 163–176. https://doi.org/10.1177/01466210122031984

Kim, S.-H., Cohen, A. S., Baker, F. B., Subkoviak, M. J., & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika*, *59*(3), 405–421. https://doi.org/10.1007/BF02296133

Levy, R., & Mislevy, R. (2016). *Bayesian psychometric modeling*. CRC Press.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Liu, Y., & Yang, J. S. (2018). Interval estimation of latent variable scores in item response theory. *Journal of Educational and Behavioral Statistics*, *43*(3), 259–285. https://doi.org/10.3102/1076998617732764

McElreath, R. (2016). *Statistical rethinking*. Taylor & Francis.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195. https://doi.org/10.1007/BF02293979

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Frontiers in Psychology*, *7*, Article 1422. https://doi.org/10.3389/fpsyg.2016.01422

Paek, I., Cui, M., Öztürk Gübes, N., & Yang, Y. (2018). Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods. *Educational and Psychological Measurement*, *78*(4), 569–599. https://doi.org/10.1177/0013164417715738

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, *22*(1), 59–73. https://doi.org/10.1214/088342307000000014

Polson, N., & Scott, J. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902. https://doi.org/10.1214/12-BA730

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: Effects of prior specifications on parameter estimates. *Behaviormetrika*, *37*(2), 87–110. https://doi.org/10.2333/bhmk.37.87

Sheng, Y. (2013). An empirical investigation of Bayesian hierarchical modeling with unidimensional IRT models. *Behaviormetrika*, *40*(1), 19–40. https://doi.org/10.2333/bhmk.40.19

Sheng, Y. (2017). Investigating a weakly informative prior for item scale hyperparameters in hierarchical 3PNO IRT models. *Frontiers in Psychology, 8*, Article 123. https://doi.org/10.3389/fpsyg.2017.00123

Stan Development Team. (2016). *RStan: The R interface to Stan* (Version 2.14.1). http://mc-stan.org/users/interfaces/rstan.html

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

Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal of Applied Measurement*, *5*(1), 48–61.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*(3), 175–191. https://doi.org/10.1007/BF02294110

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*(1), 27–51. https://doi.org/10.1177/0146621602239475

Turner, B. M., Sederberg, P. M., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384. https://doi.org/10.1037/a0032222

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*(2), 264–290. https://doi.org/10.1177/0013164411410056