*Article*

# Benefits of the Curious Behavior of Bayesian Hierarchical Item Response Theory Models—An in-Depth Investigation and Bias Correction

Applied Psychological Measurement
2024, Vol. 48(1-2) 38–56

DOI: 10.1177/01466216241227547
journals.sagepub.com/home/apm


**Christoph König**[1] **and Rainer W. Alexandrowicz**[2]

## Abstract

When using Bayesian hierarchical modeling, a popular approach for Item Response Theory (IRT) models, researchers typically face a tradeoff between the precision and accuracy of the item parameter estimates. Given the pooling principle and variance-dependent shrinkage, the expected behavior of Bayesian hierarchical IRT models is to deliver more precise but biased item parameter estimates, compared to those obtained in nonhierarchical models. Previous research, however, points out the possibility that, in the context of the two-parameter logistic IRT model, the aforementioned tradeoff has not to be made. With a comprehensive simulation study, we provide an in-depth investigation into this possibility. The results show a superior performance, in terms of bias, *RMSE* and precision, of the hierarchical specifications compared to the nonhierarchical counterpart. Under certain conditions, the bias in the item parameter estimates is independent of the bias in the variance components. Moreover, we provide a bias correction procedure for item discrimination parameter estimates. In sum, we show that IRT models create a unique situation where the Bayesian hierarchical approach indeed yields parameter estimates that are not only more precise, but also more accurate, compared to nonhierarchical approaches. We discuss this beneficial behavior from both theoretical and applied point of views.

## Keywords

Item response theory, psychometrics, bayesian hierarchical modeling, bias correction, bayesian Item response theory, weakly informative prior

[1]Goethe University Frankfurt, Germany
[2]University of Klagenfurt, Austria

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

Bayesian hierarchical modeling is a popular approach for Item Response Theory (IRT) models. They are quite complex and, depending on sample size and test length, consist of many parameters of the same type (e.g., item discriminations and item difficulties, as well as person parameters), making them excellent candidates for Bayesian hierarchical specifications. Due to their hierarchical prior structure and the associated pooling process, which maximizes the information in a given dataset, Bayesian hierarchical models yield item parameter estimates that are more precise than those of their nonhierarchical counterparts are (e.g., Katahira, 2016). This is typically reflected by narrower 95% highest density intervals (HDI) of the parameter estimates.

There is a tradeoff, however, because the increased precision (i.e., smaller standard error) is associated with a decreased accuracy (i.e., larger bias) of the parameter estimates. The pooling process depends on the variance of the individual item parameter estimates. To the extent their variance decreases, their estimates shrink towards their grand mean, that is, the mean of their hyperprior distribution (Efron & Morris, 1977). Thus, since individual item parameters always vary to some degree (Fox, 2010), their estimates obtained with Bayesian hierarchical models exhibit a certain amount of bias, proportional to the amount of shrinkage. Hence, the expected (typical) behavior of Bayesian hierarchical models is to deliver more precise but biased individual parameter estimates, compared to the parameter estimates obtained with their nonhierarchical counterparts.

However, Koenig et al. (2020) found that their optimized hierarchical two-parameter logistic (OH2PL) IRT model for small-sample item calibration outperformed its nonhierarchical counterpart, especially in terms of bias of the item discrimination parameters. This was an interesting finding, because it contradicts the typical and theoretically expected behavior of Bayesian hierarchical models (larger bias of all parameters compared to nonhierarchical models).

It is possible, however, that applying the Bayesian hierarchical approach to IRT models creates a unique situation where there is no tradeoff between accuracy and precision. Reasons for this unique situation may relate to a combination of characteristics of the item parameters of IRT models with the current practice of Bayesian hierarchical modeling (i.e., current recommendations for model specifications and the specification of priors for variance components). Koenig et al. (2020) did not investigate this possibility further. Therefore, the objective of this paper is to investigate the question whether Bayesian hierarchical IRT models indeed behave differently than their general counterparts, in the sense that the aforementioned tradeoff between precision and accuracy does not have to be made in general, or whether the behavior is a consequence of the interplay between item parameter and model characteristics when applying the Bayesian hierarchical approach to IRT models. We further want to explore the specific reasons for this atypical, but beneficial behavior of Bayesian hierarchical IRT models.

In the following sections, we illustrate (1) the core characteristics and specification of the Bayesian H2PL, (2) the typical characteristics of parameters in IRT contexts and priors of current Bayesian hierarchical IRT models, and (3) describe our comprehensive simulation study. We then (4) present the results of our simulation, and discuss them in relation to their benefits for accurate item calibration in small samples, computerized adaptive testing (CAT) and Bayesian hierarchical IRT modeling in general. Scripts to replicate this simulation, our data, and results are available as an online supplement at https://osf.io/ybk2f/ (Jackman, 2009)

## Pooling, Shrinkage, and Bias in the Context of the Hierarchical 2PL Item Response Theory Model

Suppose a sample of $J$ ($j = 1, \ldots, J$) individuals takes a test consisting of $K$ ($k = 1, \ldots, K$) items. According to the 2PL IRT model (Birnbaum, 1968), the probability of a correct response

$y_{kj} \in \{0, 1\}$ of person $j$ to item $k$ is a function of their ability $\theta_j$, the discrimination $\alpha_k$, and difficulty $\beta_k$ of the item

$$\Pr(y_{jk} = 1|\theta_j, \alpha_k, \beta_k) = \text{logit}[\alpha_k (\theta_j - \beta_k)]. \quad (1)$$

Both discrimination and difficulty parameters are item-specific. Thus, in a test of $K$ items there are $K$ discrimination and difficulty parameters. The hierarchical Bayesian approach implements a hierarchical structure of prior distributions for the individual item parameters $\alpha_k, \beta_k$ and their grand means $\mu_\alpha, \mu_\beta$ along with the variance components $\tau_\alpha, \tau_\beta$ (cf. Koenig et al., 2020).

A common implementation of the hierarchical 2PL IRT model is as follows. For the abilities $\theta_j$, (a) standard normal distribution is used as prior. Moreover, for each item, a bivariate normal distribution is used for the item-specific parameter vector $\xi_k = \{\log(\alpha_k), \beta_k\}$. The log-transformation on the discrimination parameter is required for the use of the bivariate normal distribution (e.g., Glas & van der Linden, 2003). The bivariate normal distribution is governed by the vector of item parameter grand means $\mu_{\alpha,\beta} = \{\mu_\alpha, \mu_\beta\}$ and covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} \tau_\alpha & \rho_{\beta\alpha}\tau_\beta\tau_\alpha \\ \rho_{\alpha\beta}\tau_\alpha\tau_\beta & \tau_\beta \end{bmatrix}$. For the grand means $\mu_{\alpha,\beta}$ normal prior distributions are used. The prior for $\Sigma$ is a noninformative Inverse Wishart distribution with $\upsilon = 3$ degrees of freedom and the identity matrix $\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ as scale matrix. The variance components $\tau_\alpha$ and $\tau_\beta$ are not modeled explicitly. They can be recovered, however, from the diagonal of the covariance matrix $\Sigma$. The full model is specified as follows:

$$\Pr(y_{ij} = 1|\theta_j, \alpha_k, \beta_k) = \text{Bernoulli}(\text{logit}[\alpha_k (\theta_j - \beta_k)]) \quad (2)$$

Level 1:

$$\theta_j \sim N(0, 1) \quad (3)$$

$$\xi_k \sim \text{BVN}(\mu_{\alpha,\beta}, \Sigma) \quad (4)$$

Level 2:

$$\mu_\alpha \sim N(0, 1) \quad (5)$$

$$\mu_\beta \sim N(0, 2) \quad (6)$$

$$\Sigma \sim \text{IW}(3, I) \quad (7)$$

In this hierarchical structure, the individual item parameters share an inherent dependency with their respective grand means (Betancourt & Girolami, 2015). This dependency maximizes the information available for the estimation of the individual item parameter estimates. The increase in information leads to an increased precision of the individual parameter estimates, that is, narrower 95% HDIs or smaller standard errors.

Another consequence of the dependency of individual item parameters and their grand mean is that, for instance, an item discrimination parameter $\alpha_k$ lies between the individual estimate $\widehat{\alpha}_k$ and its grand mean $\mu_\alpha$. Thus, there are two extreme cases: (1) the item discrimination parameter estimate corresponds to its individual estimate $\alpha_k = \widehat{\alpha}_k$ (no pooling), or (2) the item discrimination parameter estimate corresponds to its grand mean $\alpha_k = \mu_\alpha$ (complete pooling). The estimates of the item discrimination parameters depend on their variance component $\tau_\alpha$. As $\tau_\alpha \to 0$, the item discrimination parameter estimates shrink towards, and eventually correspond to, their grand

mean $\mu_\alpha$. This shrinkage, however, introduces a certain amount of bias into the individual item parameter estimates, because $\alpha_k = \mu_\alpha$ would only be unbiased if $\tau_\alpha = 0$, that is, all item discrimination parameter estimates are equal. In the context of the 2PL model, however, zero variance situations are hardly realistic. Thus, in the case of $\tau_\alpha > 0$ and the resulting shrinkage, individual parameter estimates in Bayesian hierarchical models, although more precise, are biased compared to estimates obtained from nonhierarchical models.

The amount of bias introduced into the estimation of the individual item parameters depends on the relation of the true variance $\tau_\alpha$ of a set of item parameters (item discriminations in this case) and the accuracy of its estimate $\widehat{\tau}_\alpha$. Suppose you have a set of item discrimination parameters, the typical range of their values is known and similar across assessments, that is, their grand mean remains roughly constant across applications (which seems a plausible assumption for item discrimination parameters of the 2PL). The theoretically possible shrinkage (and thus the associated bias) increases with an increasing $\tau_\alpha$, because with the increasing $\tau_\alpha$ the individual parameter estimates are less tightly clustered around their grand mean. Consequently, the individual parameter estimates' biases increase even more with the bias of $\widehat{\tau}_\alpha$. Their bias should be larger when $\widehat{\tau}_\alpha$ is underestimated, because less variability is assumed than there actually is, leading to an increased amount of unintended shrinkage. Thus, we have a relation between $\widehat{\tau}_\alpha$ the accuracy of and the respective individual item discrimination parameter estimates $\widehat{\alpha}_k$. We will therefore explore in detail the relation of the bias of $\widehat{\tau}_\alpha$ and $\widehat{\alpha}_k$.

## Typical Characteristics of Parameters and Current Specifications of Bayesian Hierarchical Item Response Theory Models

To derive possible explanations of the atypically better performance of the Bayesian H2PL model compared to its nonhierarchical counterpart (as noted by Koenig et al., 2020), we have to consider the typical characteristics of the item parameters, as well as processes in the context of the current practice of specifying Bayesian hierarchical IRT models.

First, the item discrimination and difficulty parameters are known to fall in a relatively narrow range. The item discriminations, for example, typically fall in the interval $[0.5, 3.0]$ (e.g., OECD, 2021). Parameter values outside this interval are considered unrealistic or undesirable, and items exhibiting such discriminations are usually eliminated from an item bank. Similarly, item difficulty parameters are typically found to be in the interval $[-4, 4]$ (e.g., OECD, 2021); again, item difficulties outside this interval are rare and seldom used for a test. Hence, the variance component of both item parameter types is also restricted. Using these assumptions, the maximum variance of the item discriminations is $\tau_\alpha = 1$, and of the item difficulties $\tau_\beta = 16$. Note that the typical variance components in calibrated item banks are much smaller: For instance, the variance components of the item discrimination and difficulty parameters of the PISA 2018 cycle were mostly $\tau_\alpha < 0.4$, and $\tau_\beta < 1.0$ (cf. OECD, 2021). Hence, both individual item parameter estimates cluster closely around their respective grand means (especially in case of the item discrimination parameters). Thus, the variance component can be considered small enough to render the bias induced by shrinkage negligible, relative to the true values of the item parameters.

Second, current specifications of Bayesian H2PL models are tailored towards avoiding bias in both the individual parameter estimates $\widehat{\alpha}_k, \widehat{\beta}_k$ and the estimates of their variance components $\widehat{\tau}_\alpha, \widehat{\tau}_\beta$. They use separate prior distributions for the variance components and the correlation between the item parameters, for a better control over the behavior of both parameters (e.g., Barnard et al., 2000), and to avoid the a-priori dependencies associated with the Inverse Wishart distribution for covariance matrices. This so-called separation strategy uses the LKJ prior distribution for the lower-triangular Cholesky factor of the correlation matrix $\mathbf{L}_\Omega$ of the item

parameters. For a $D \times D$ lower-triangular Cholesky factor of a positive-definite matrix this prior distribution has the density $\text{LKJ}(\mathbf{L}_\Omega|\eta) = \prod_{d=2}^{D} \text{L}_{dd}^{D-d+2\eta-2}$ governed by the shape parameter $\eta$ (Lewandowski et al., 2009). Setting $\eta = 1$ yields a uniform density over correlation matrices of order $D$; as $\eta$ increases, extreme correlations become more unlikely (Stan Development Team, 2020). Setting $\eta = 2$ yields a weakly informative prior for the Cholesky factor of the item correlation matrix. For the variance components, instead of the noninformative Inverse Gamma distribution, current Bayesian hierarchical models use weakly informative (half-) Cauchy or Exponential distributions as prior distributions for the variance components (e.g., Ulitzsch et al., 2020; Bezirhan et al., 2021). The primary reason for this shift is that the Inverse Gamma distribution is problematic for variance components close to zero. Especially in noninformative specifications, the Inverse Gamma distribution has a low mass near zero, which introduces unintended information (Gelman, 2006). When the true variance component is near zero (in contrast to the assumed prior during estimation), its estimate will be drawn away from zero, resulting in biased estimates of the variance components. Using either the Cauchy or Exponential distribution eliminates (or, at least, reduces) this source of bias. Thus, both alternative prior distributions yield more accurate variance estimates. In the model specification below, the separation strategy is implemented by (13) and (14).

Lastly, current specifications of Bayesian H2PL models are noncentered. In noncentered specifications, the first level of the model consists of a standard normal prior distribution for the abilities $\theta_j$. But instead of directly sampling item-specific parameter vectors $\xi_k$ from a bivariate normal distribution as in (4), the first level of the noncentered model is completed by a standard normal prior distribution for item-specific vectors of uncorrelated z-scores $\tilde{\xi}_k = \{z_{\alpha_k}, z_{\beta_k}\}$ (see also Koenig et al., 2020, 2022). These z-scores are essentially uncorrelated deviations from the parameter grand means $\mu_\alpha$ and $\mu_\beta$. The actual item parameter estimates $\widehat{\alpha}_k, \widehat{\beta}_k$ are then computed, not sampled, by first multiplying $\tilde{\xi}_k$ by the diagonal matrix of the variance components $\Lambda$ and $\mathbf{L}_\Omega$ to obtain item-specific vectors of correlated deviations, $\xi_k = (\Lambda \mathbf{L}_\Omega \tilde{\xi}_k)^{\text{T}}$, and then adding the respective grand means, that is, $\widehat{\alpha}_k = \exp(\mu_\alpha + \xi_{\alpha_k})$ and $\widehat{\beta}_k = \mu_\beta + \xi_{\beta_k}$ (e.g., Koenig et al., 2022). The resulting joint posterior of the item and person parameters is much easier to explore [no correlation on level 1 and no cross-level dependency as in (4)], thus avoiding bias due to inefficient sampling or an insufficient effective sample size (Betancourt & Girolami, 2015; Zitzmann & Hecht, 2019). The complete noncentered, optimized specification of the H2PL is as follows:

$$\Pr(y_{jk} = 1|\theta_j, \alpha_k, \beta_k) = \text{Bernoulli}\left(\text{logit}\left[\alpha_k \left(\theta_j - \beta_k\right)\right]\right) \tag{8}$$

Level 1:

$$\theta_j \sim \text{N}(0, 1) \tag{9}$$

$$\tilde{\xi}_k \sim \text{N}(0, 1) \tag{10}$$

Level 2:

$$\mu_\alpha \sim \text{N}(0, 1) \tag{11}$$

$$\mu_\beta \sim \text{N}(0, 2) \tag{12}$$

$$\tau_{\alpha,\beta} \sim \text{Cauchy}(0, 5) \tag{13}$$

$$L_\Omega \sim \text{LKJ}(2) \tag{14}$$

In sum, favorable and optimized model specifications that avoid bias in estimated variance components $\widehat{\tau}_\alpha, \widehat{\tau}_\beta$ in combination with typically small true variance components $\tau_\alpha, \tau_\beta$ of the item parameters might explain why the Bayesian H2PL yields item parameter estimates that are less biased than the estimates obtained with its nonhierarchical counterpart. We will now explore this principle further.

## Purpose of the Study

Consequently, the primary purpose of this paper is an in-depth investigation of the question, whether the curious behavior of the Bayesian H2PL is (a) an indication of a generally different behavior of Bayesian hierarchical IRT models, compared to that of their general counterparts, or (b) a consequence of the interplay between item parameter and model characteristics as outlined above. Moreover, we aim at providing insights regarding the specific reasons for this curious behavior in the context of the Bayesian H2PL.

Therefore, we follow a two-step approach answering two primary research questions. First, we investigate whether the hierarchical specifications of the 2PL, namely, the optimized Bayesian hierarchical 2PL (OH2PL; Koenig et al., 2020) and the standard Inverse Wishart specification (SH2PL), yield less biased item parameter estimates than their nonhierarchical counterpart. We chose the OH2PL and the SH2PL as examples of current approaches to Bayesian hierarchical IRT modeling (e.g., Gilholm et al., 2021). In this step, we compare the performance (relative and absolute bias, root mean squared error *RMSE*) of the hierarchical specifications of the 2PL with different specifications of the nonhierarchical 2PL model to check whether the advantages are robust across a broad range of data conditions. We also look at the widths of the 95% HDIs of the resulting item parameter estimates across model specification to assess the precision of the estimates. Second, we take a closer look at the relation of the relative and the absolute bias in the individual parameter estimates $\widehat{\alpha}_k, \widehat{\beta}_k$ on the one hand, and the true and estimated variance components $\tau_\alpha, \tau_\beta$ in the hierarchical specifications of the 2PL on the other hand. Here, we aim at clarifying (1) whether the relative and absolute bias in the item parameter estimates is independent from their true variance components (i.e., does not increase when $\tau_\alpha$ increases), and (2) whether the relative and absolute bias in the item parameter estimates is independent from the bias in the estimated variance components. If both questions can be answered in the affirmative, we provide evidence that the behavior of the H2PL is in fact different from (or even superior to) general Bayesian hierarchical models.

As a further contribution to the literature, we will further seek clarification whether there is a critical value of the true variance components $\tau_{\text{Critical}}$ that can be considered too large, that is, a cutoff from which we have to expect biased individual item parameters. This, in turn, will allow pinpointing cases in which the Inverse Gamma is a better choice for the hyperprior for the variance components. As mentioned before, the use of the Inverse Gamma has been discouraged because of its erratic behavior when the true variance is close to zero (e.g., Gelman, 2006; Polson & Scott, 2012). So far, there is no clear indication of what might be considered "too close to zero."

Moreover, we present a bias correction procedure for individual item discrimination parameter estimates in cases in which they are biased because of their variance component being underestimated. As mentioned before, bias should be more pronounced when the true variance is underestimated due to larger unintended shrinkage. Such a procedure constitutes another important contribution to optimize the Bayesian H2PL model further, especially for its use in small-sample situations.

## Method

### Simulation Design

The fully crossed design of the study consisted of the following factors. (1) The variance in the item discriminations $\tau_\alpha = (0.05, 0.1, 0.2, 0.4, 0.6, 0.75, 1.0)$, (2) the variance in the item difficulties $\tau_\beta = (0.45, 0.6, 0.8, 1.0, 1.5)$, (3) the correlation between the item parameters $\rho_{\alpha\beta} = (.0, .3)$, (4) the sample size $N = (50, 100, 200, 500, 1200)$, (5) the test length $K = (20, 30, 40)$, and (6) the model specification (the OH2PL, the SH2PL, and three specifications of the nonhierarchical 2PL model (NH2PL): noninformative, weakly informative, and informative prior distribution for the item discrimination parameters; see below). This resulted in a total of 5250 conditions examined. We chose the variances of the item parameters to cover a large range of both typical and extreme values and the correlations, sample sizes, and test lengths to be able to investigate, whether the beneficial behavior of the OH2PL found by Koenig et al. (2020) is also present when the item parameters are uncorrelated, and in suboptimal testing conditions. Sample sizes larger than $N = 200$ are considered adequate for item calibration under the 2PL IRT model, where $N = 1200$ is considered the point of diminishing returns (De Ayala, 2023).

### Nonhierarchical Specifications of the 2PL Model

The prior configurations of the nonhierarchical models were chosen to keep the different model specifications comparable, and reflect prior configurations common in IRT modeling (e.g., Levy & Mislevy, 2016). In all model specifications, the ability parameters were given a standard normal prior $\theta_j \sim N(0, 1)$ for identification purposes.

The nonhierarchical Bayesian specifications of the 2PL model only have a single level consisting of prior distributions for the individual item parameters. Because Koenig et al. (2020) found differences in the performance (compared to the hierarchical specification) to be specific to the item discrimination parameter, the specifications differ primarily in the prior distribution for the individual discrimination parameters [NH2PL1, NH2PL2, NH2PL3, respectively: $\alpha_k \sim \log N(0, 1), \alpha_k \sim \log N(0.5, 1), \alpha_k \sim \log N(1, 1)$]. The prior distribution for the item difficulty parameters $\beta_k \sim N(0, 2)$ was kept constant across the three specifications.

### Data Generation and Analysis

Data were generated under a unidimensional 2PL with correlated item parameters. To obtain realistic item discrimination and difficulty parameters ($0.5 < \alpha_k < 4.0$ and $-4 < \beta_k < 4$), we drew uncorrelated vectors of item parameters from a truncated bivariate normal distribution with grand mean vector $\mu_{\alpha,\beta} = \{1, 0\}$ with lower limits $LL_{\alpha,\beta} = \{0.65, -4.5\}$ and upper limits $UL_{\alpha,\beta} = \{4.0, 4.5\}$. We rescaled these vectors by mean centering the uncorrelated vectors and adding the true marginal means of the truncated bivariate distribution to obtain the desired correlations. We generated 100 data sets for each simulation condition with different item and person parameters for each dataset resulting in slightly more than half a million data sets.

We used Stan (Carpenter et al., 2017) and the *R* interface *Rstan* (Stan Development Team, 2020) to estimate the hierarchical and nonhierarchical models. Stan employs the No-U-Turn-Sampler (NUTS; Hoffman & Gelman, 2014), which is an adaptive variant of Hamiltonian Monte Carlo (HMC). In HMC, Hamiltonian systems are simulated to sample from target distributions (Neal, 2011). By introducing the momentum as an auxiliary variable, HMC is able to utilize the local geometry of the target distribution in order to traverse the posterior density more efficiently (Gelman et al., 2014). This usually requires, however, hand-tuning of key parameters of the

standard HMC algorithm. The No-U-Turn-Sampler implemented in Stan eliminates this requirement by adaptively tuning the necessary parameters. Thus, applied researchers can focus on the model specification, and not on the setup of the MCMC algorithm (Annis et al., 2017). For more details about the standard HMC algorithm and its adaptive variant interested readers are referred to Hoffman and Gelman (2014), where both algorithms are illustrated in great detail. Three chains with 3000 draws (1000 burn-in cycles) were set up. Moreover, different random starting values were supplied to each chain. Convergence was achieved when the R-hat diagnostic was smaller than 1.05 (Vehtari et al., 2021). For a comprehensive overview of the frequency of non-convergent solutions across model specifications, see Supplement 1. Non-convergent solutions were excluded from further analysis.

### Evaluation Criteria

To test the aforementioned assumptions, we calculated the average raw bias $\left(B = \frac{1}{R}\pi_{est} - \pi_{true}\right)$, the average relative bias [because item difficulties can be close to zero, we calculated it as $B_{\text{rel}} = \frac{1}{R}(\pi_{\text{est}} - \pi_{\text{true}})/(1 + |\pi_{\text{true}}|)$], the absolute bias $B_{abs} = |\frac{1}{R}\pi_{\text{est}} - \pi_{\text{true}}|$, and the root mean squared error ($RMSE = \sqrt{\sum_R(\pi_{est} - \pi_{true})^2/R}$). Here, $\pi_{\text{est}}$ and $\pi_{\text{true}}$ are the estimated and true values of a parameter ($\pi$ serving as generic notation for the parameters of interest), respectively, and $R$ the number of replications. The average width of the 95% HDI was indicated by the difference between the 97.5% percentile and the 2.5% percentile, and was averaged over items and replications. As mentioned above, the behavior of the hierarchical 2PL IRT models is in fact different from (or even superior to) general Bayesian hierarchical models when two assumptions are met. First, the average relative and absolute bias and *RMSE* of the item parameters under the hierarchical approach is similar or superior to the relative and absolute bias and *RMSE* of the item parameters under its nonhierarchical counterpart. The first assumption will be supported, if $B_{\text{H2PL}} \leq B_{\text{2PL}}$, $B_{\text{absL}_{\text{H2PL}}} \leq B_{\text{abs}_{\text{2PL}}}$ and $RMSE_{\text{H2PL}} \leq RMSE_{\text{2PL}}$ across a wide range of simulation conditions. Second, a different behavior is assumed when the relative and absolute bias in the item parameter estimates does not increase when the associated true variance components increase, and when the bias in the item parameter estimates is independent from the bias in the estimated variance components. The second assumption will be supported if $B_{\text{rel}}$ remains approximately constant as $\tau_\alpha, \tau_\beta$ increase and $r_{b_{\hat{\alpha}_k} b_{\hat{\tau}_\alpha}} = r_{b_{\hat{\beta}_k} b_{\hat{\tau}_\beta}} = r_{ab_{\hat{\alpha}_k} b_{\hat{\tau}_\alpha}} = r_{ab_{\hat{\beta}_k} b_{\hat{\tau}_\beta}} \approx 0$. We used the commonly applied cutoff of $B_{\text{rel}} < 0.10$ to indicate unbiased item parameter estimates (Kaplan, 1988). The second indication was independence of $B_{\text{rel}}$ and $B_{\text{abs}}$ in individual item parameter estimates and their variance components.

## Results

### Hierarchical Specifications Consistently Outperform the Nonhierarchical 2PL

Figure 1 shows the bias of the item discrimination parameter estimates of both specifications of the Bayesian H2PL in comparison with their nonhierarchical counterpart, across all simulation conditions. It becomes evident that the performance of both the OH2PL and the SH2PL was better (and never worse) than the performance of the different specifications of the nonhierarchical 2PL. This general pattern also held when investigating the absolute bias of the item discrimination parameters (Figure 2). The advantages of the hierarchical specifications were especially pronounced with $N = 50$ observations. Both hierarchical specifications outperforming the nonhierarchical 2PL was corroborated by the differences in

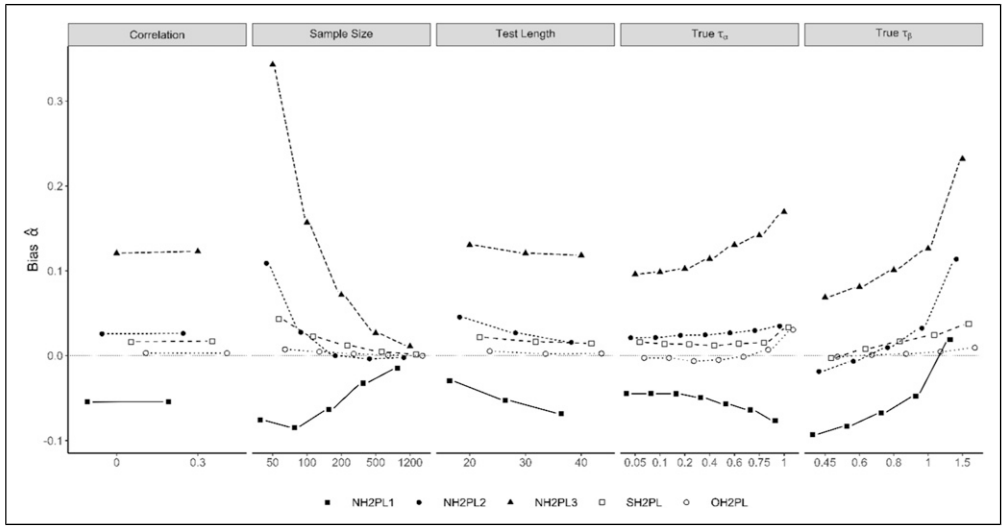


**Figure 1.** Raw bias in item discrimination parameter estimates across simulation conditions.

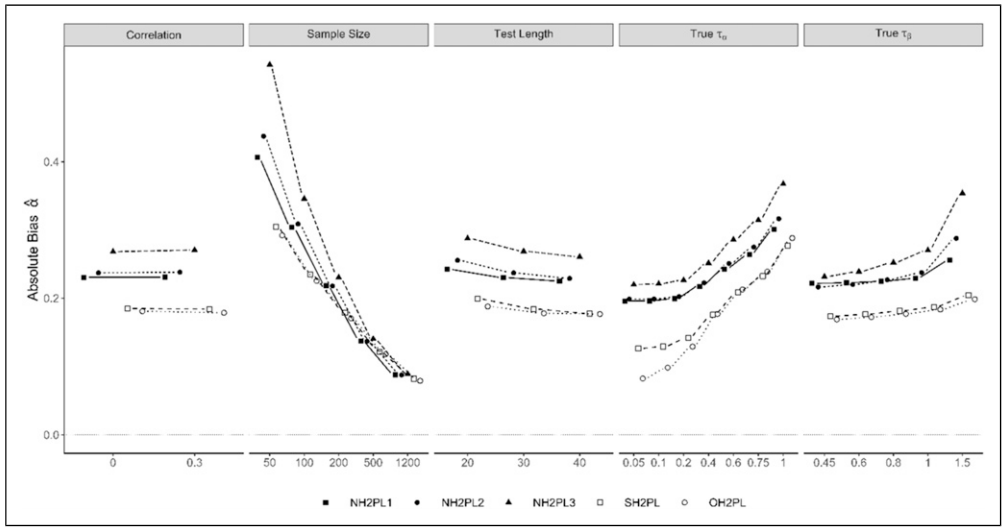


**Figure 2.** Absolute bias in item discrimination parameter estimates across simulation conditions.

the *RMSE* of the item discrimination parameters (see Figure 3). The OH2PL and the SH2PL consistently outperformed their nonhierarchical counterpart. As expected, the average width of the 95% HDI was consistently smaller in the hierarchical specifications, compared to their nonhierarchical counterpart (Figure 4). In sum, the first condition in favor of an atypical behavior of Bayesian hierarchical IRT models was met: the hierarchical specifications yielded indeed parameter estimates that were less biased compared to their nonhierarchical counterpart. Moreover, they were robust (i.e., they did not depend on the specification of the nonhierarchical 2PL), and they were not specific to the OH2PL, although the OH2PL

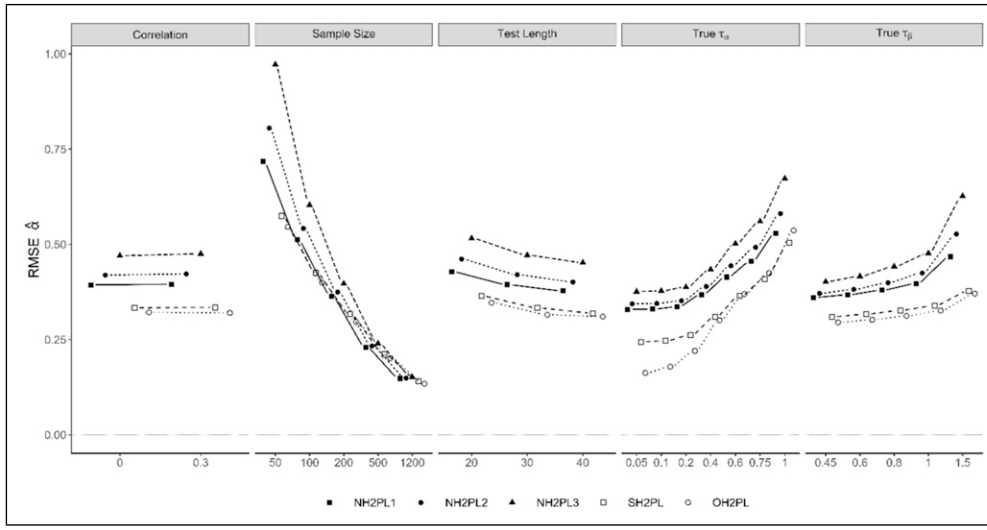


**Figure 3.** *RMSE* in item discrimination parameter estimates across simulation conditions.

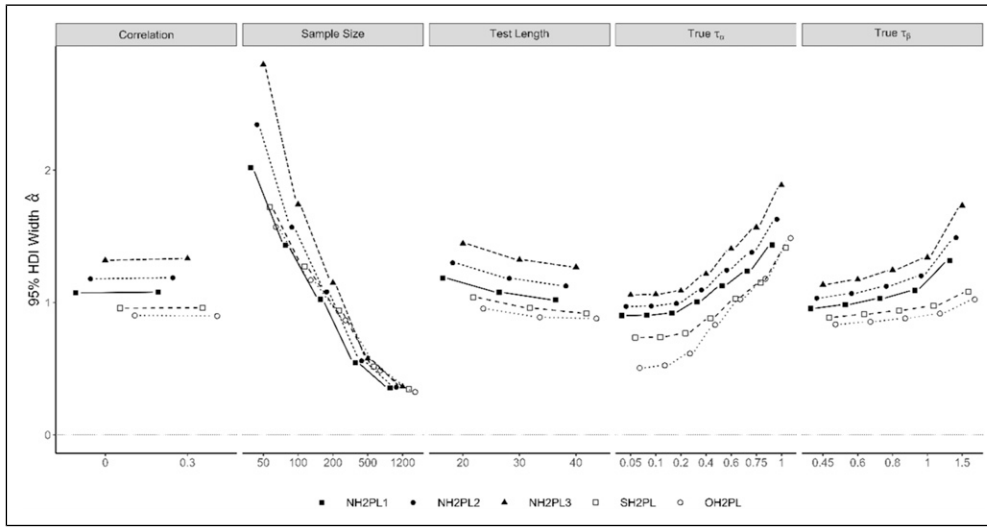


**Figure 4.** 95% HDI of the item discrimination parameter estimates across simulation conditions.

(slightly) outperformed the SH2PL. Overall, the results for the item difficulty parameter were similar and did not lead to different conclusions. They are included in Supplement 1. The general pattern of advantages of the hierarchical specifications, especially in smaller samples, could also be found in models that are more complex, such as the generalized partial credit model (GPCM; Muraki, 1992). We ran a small additional simulation with this model that corroborated the findings for the 2PL model. The additional simulation is described in Supplement 3.

### Bias in Item Parameters Partly Independent from Bias in Variance Components

Table 1 summarizes the relation between the relative and absolute bias in the item parameter estimates $\widehat{\alpha}_k, \widehat{\beta}_k$ and the relative bias in the estimates of their associated variance components $\widehat{\tau}_\alpha, \widehat{\tau}_\beta$. There was no association between the relative bias in both item parameter estimates and the relative bias in their variance components (the correlations were negligible). In contrast, the correlations between the absolute bias in both item parameter estimates and the relative bias in the estimates of their associated variance components were not negligible.

Hence, the second condition supporting an atypical behavior of Bayesian hierarchical IRT models (namely, $B_{\text{rel}}$ remaining approximately constant as $\tau_\alpha, \tau_\beta$ increase and $r_{ab_{\hat{\alpha}_k} b_{\hat{\tau}_\alpha}} = r_{ab_{\hat{\beta}_k} b_{\hat{\tau}_\beta}} \approx 0$) was only partly met. Analogously to the first condition, this applied to both the OH2PL and the SH2PL.

### Bias in Individual Item Parameter Estimates and the True Variance Components

The violin plots in Figure 5 illustrate the change in relative bias in the individual item parameter estimates along increasing true values of the associated variance components $\tau_\alpha, \tau_\beta$. The violin plots in Figure 5 show the means and $\pm 2SD$ (black dots with associated vertical bars) of the relative bias in $\alpha$, along with the kernel probability density of the relative bias at the true values of the variance components.

From Figure 5 (right panel) we learn that the relative bias in the item difficulty parameter $\beta$ did not increase with the true variance $\tau_\beta$. The same applied to the item discrimination parameter $\alpha$ (left panel), but for $\tau_\alpha$ exceeding .4, we observed an increasing amount of outliers larger than .1. However, all instances the middle 50% of estimates were still within the interval [–.1, .1] (dashed lines in Figure 5).

**Table 1.** Correlations Between the Bias in the Estimates of the Variance Components and the Relative and Absolute Bias in the Item Parameter Estimates Across Sample Sizes and Test Lengths.

| | | *OH2PL* | | | | *SH2PL* | | | |
|---|---|---|---|---|---|---|---|---|---|
| *N* | *K* | $r_{b_{\hat{\alpha}_k} b_{\hat{\tau}_\alpha}}$ | $r_{b_{\hat{\beta}_k} b_{\hat{\tau}_\beta}}$ | $r_{ab_{\hat{\alpha}_k} b_{\hat{\tau}_\alpha}}$ | $r_{ab_{\hat{\beta}_k} b_{\hat{\tau}_\beta}}$ | $r_{b_{\hat{\alpha}_k} b_{\hat{\tau}_\alpha}}$ | $r_{b_{\hat{\beta}_k} b_{\hat{\tau}_\beta}}$ | $r_{ab_{\hat{\alpha}_k} b_{\hat{\tau}_\alpha}}$ | $r_{ab_{\hat{\alpha}_k} b_{\hat{\tau}_\beta}}$ |
| 50 | 20 | .070 | −.005 | −.115 | −.202 | .008 | −.013 | .215 | .06 |
| | 30 | .043 | .001 | −.218 | −.261 | .006 | −.005 | .169 | .05 |
| | 40 | .024 | .003 | −.255 | −.286 | .006 | −.002 | .167 | .04 |
| 100 | 20 | .026 | −.003 | −.200 | −.218 | −.004 | −.004 | .181 | .04 |
| | 30 | .005 | −.009 | −.259 | −.257 | −.003 | −.008 | .150 | .04 |
| | 40 | .004 | −.007 | −.278 | −.282 | .001 | −.007 | .118 | .03 |
| 200 | 20 | .008 | −.011 | −.233 | −.218 | −.005 | −.011 | .160 | .02 |
| | 30 | −.001 | −.001 | −.267 | −.247 | −.012 | −.002 | .128 | .03 |
| | 40 | −.009 | .001 | −.266 | −.259 | −.010 | −.011 | .114 | .01 |
| 500 | 20 | .001 | .003 | −.223 | −.194 | −.009 | −.019 | .153 | .02 |
| | 30 | −.009 | −.009 | −.234 | −.213 | −.005 | −.010 | .120 | .01 |
| | 40 | −.010 | −.005 | −.227 | −.224 | −.006 | −.008 | .109 | .01 |
| 1200 | 20 | −.003 | −.006 | −.201 | −.174 | −.017 | −.017 | .147 | .06 |
| | 30 | −.001 | −.007 | −.208 | −.186 | −.010 | −.002 | .129 | .03 |
| | 40 | −.012 | −.010 | −.213 | −.189 | −.001 | −.011 | .107 | .02 |

A potential reason for this increase can be found when looking at the bias in the variance component $\widehat{\tau}_\alpha$. Figure 6 reveals a clear negative relationship of true and estimated $\tau_\alpha$, in that $\widehat{\tau}_\alpha$ was *over*estimated for true $\tau_\alpha$ below .4 and *under*estimated for true $\tau_\alpha$ larger than .4. Moreover, there was a clear difference between the two model variants: While the SHPL showed a

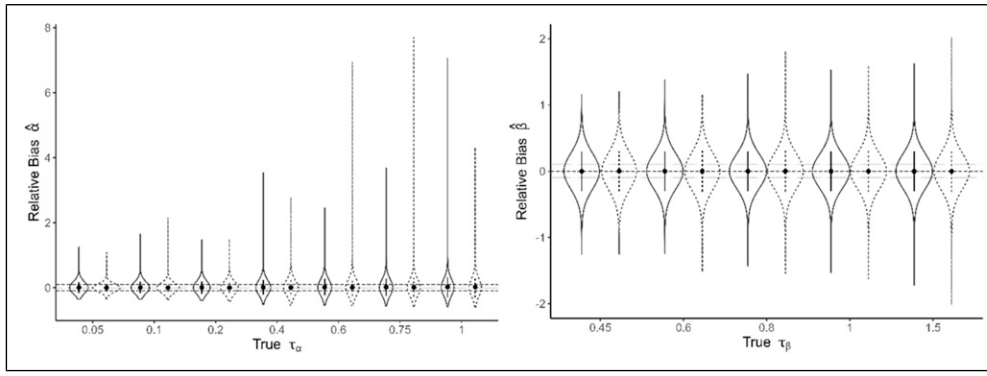


**Figure 5.** Relative Bias in the Item Parameter Estimates by True Variance Components.
Note. Left panel: Item discrimination parameters. Right Panel: Item difficulty parameters. OH2PL with solid lines, SHPL with dotted lines. The dashed lines indicate the interval [–.1, +.1].

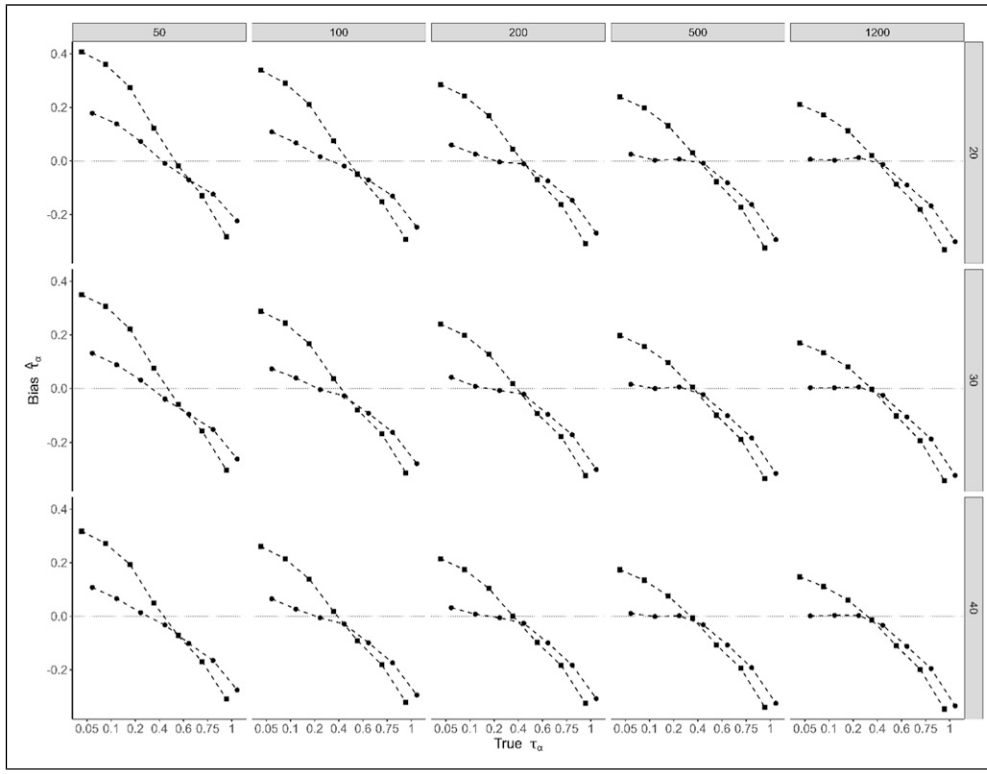


**Figure 6.** Bias of Estimated Variance Component $\widehat{\tau}_\alpha$ by its True Values $\tau_\alpha$ for Samples and Test Lengths.
Note. Circles: OH2PL. Boxes: SHPL. The pattern does not change across simulation conditions.

marked overestimation for $\tau_\alpha < .4$ and underestimation for $\tau_\alpha > .4$, only the latter applied for the OH2PL (i.e., there is just a minor overestimation for true $\tau_\alpha < .4$ but still underestimation for true $\tau_\alpha > .4$).

To investigate this potential explanation further, we ran two four-way ANOVAs with the relative and absolute bias as dependent variables and specification, sample size, test length, and true variance component $\tau_\alpha$ as independent factors. The four-way interaction between the factors was significant for both relative and absolute bias (see Section 4 in Supplement 1).

We found that the relative and absolute bias was consistently higher when $\tau_\alpha > .4$ compared to $\tau_\alpha < .4$ (see Table 2). Moreover, the largest amounts of bias were clustered in conditions with very small sample sizes, short test lengths, and large/extreme variances.

**Table 2.** Mean Bias Across Specification, Sample Size, Test Length, and Variance Components.

| | | Small | | Typical | | Large | | Extreme | |
|---|---|---|---|---|---|---|---|---|---|
| *N* | *K* | RB | AB | RB | AB | RB | AB | RB | AB |
| *OH2PL* | | | | | | | | | |
| 50 | 20 | −.002 | .159 | .005 | .251 | .024 | .390 | .049 | .506 |
| | 30 | −.002 | .138 | .010 | .235 | .023 | .354 | .037 | .450 |
| | 40 | −.010 | .129 | .010 | .231 | .019 | .337 | .035 | .427 |
| 100 | 20 | −.002 | .116 | .004 | .203 | .015 | .287 | .029 | .367 |
| | 30 | .001 | .105 | .005 | .192 | .013 | .270 | .022 | .333 |
| | 40 | .001 | .097 | .006 | .188 | .012 | .263 | .021 | .320 |
| 200 | 20 | −.001 | .086 | .003 | .162 | .010 | .215 | .014 | .259 |
| | 30 | −.001 | .081 | .004 | .154 | .010 | .199 | .011 | .243 |
| | 40 | .001 | .079 | .004 | .150 | .009 | .195 | .012 | .236 |
| 500 | 20 | −.001 | .068 | .003 | .115 | .003 | .142 | .005 | .167 |
| | 30 | .001 | .064 | .003 | .109 | .003 | .134 | .005 | .158 |
| | 40 | .001 | .063 | .003 | .106 | .004 | .130 | .005 | .153 |
| 1200 | 20 | .001 | .054 | .001 | .078 | .001 | .092 | .002 | .111 |
| | 30 | .001 | .052 | .001 | .072 | .001 | .086 | .002 | .104 |
| | 40 | .001 | .051 | .001 | .072 | .001 | .086 | .002 | .101 |
| *SH2PL* | | | | | | | | | |
| 50 | 20 | .023 | .228 | .030 | .278 | .044 | .378 | .057 | .475 |
| | 30 | .020 | .202 | .027 | .253 | .037 | .350 | .047 | .436 |
| | 40 | .019 | .185 | .024 | .241 | .038 | .345 | .042 | .422 |
| 100 | 20 | .012 | .178 | .017 | .218 | .023 | .286 | .032 | .353 |
| | 30 | .011 | .158 | .014 | .201 | .019 | .267 | .028 | .329 |
| | 40 | .011 | .147 | .015 | .195 | .018 | .261 | .024 | .318 |
| 200 | 20 | .006 | .138 | .010 | .168 | .015 | .214 | .017 | .258 |
| | 30 | .005 | .125 | .007 | .157 | .013 | .199 | .016 | .243 |
| | 40 | .006 | .118 | .008 | .152 | .012 | .196 | .014 | .236 |
| 500 | 20 | .002 | .099 | .004 | .115 | .006 | .141 | .007 | .196 |
| | 30 | .002 | .091 | .004 | .110 | .005 | .133 | .006 | .160 |
| | 40 | .002 | .087 | .004 | .106 | .005 | .130 | .006 | .154 |
| 1200 | 20 | .001 | .070 | .001 | .079 | .002 | .092 | .004 | .110 |
| | 30 | .001 | .070 | .001 | .074 | .002 | .088 | .003 | .104 |
| | 40 | .001 | .063 | .002 | .072 | .002 | .086 | .002 | .101 |

*Note.* N = Sample Size. K = Test Length; Small = $\tau_\alpha < 0.1$. Typical = $\tau_\alpha < 0.4$. Large = $\tau_\alpha < 0.6$. Extreme = $\tau_\alpha > 0.6$. All standard errors were smaller than .001; thus, they are not shown for readability. RB = Relative Bias. AB = Absolute Bias.

Thus, we may summarize that the third condition for confirming an atypical behavior of Bayesian hierarchical IRT models was also only partially fulfilled. This applied to both hierarchical specifications of the 2PL.

## A Bias Correction Procedure

Interestingly, we found a relationship between the bias in the item discrimination parameter estimates and the bias in the variance component $\widehat{\tau}_\beta$, which was unexpected. The top row of Figure 7 illustrates this relationship with a series of boxplots, summarizing the bias in $\widehat{\alpha}_k$ (the y-axis) at fixed cutpoints of the bias in $\widehat{\tau}_\beta$ (the x-axis). The dashed line over the boxplots illustrates the regression of $\widehat{\alpha}_k$ on $\widehat{\tau}_\beta$. The relationship was negative for both the SH2PL ($\beta = -0.74, SE = 0.002$) and OH2PL ($\beta = -0.52, SE = 0.002$); as the bias in the variance component increased, the bias in the item discrimination parameter estimates decreased. In other words, when the variance component was underestimated, the individual parameter estimates were more biased. This relationship allowed developing a bias-correcting procedure of the individual item discrimination parameter estimates $\widehat{\alpha}_k$ due to the bias of the estimated item difficulty variance component $\widehat{\tau}_\beta$. This correction served two purposes: (1) it made the bias in the item

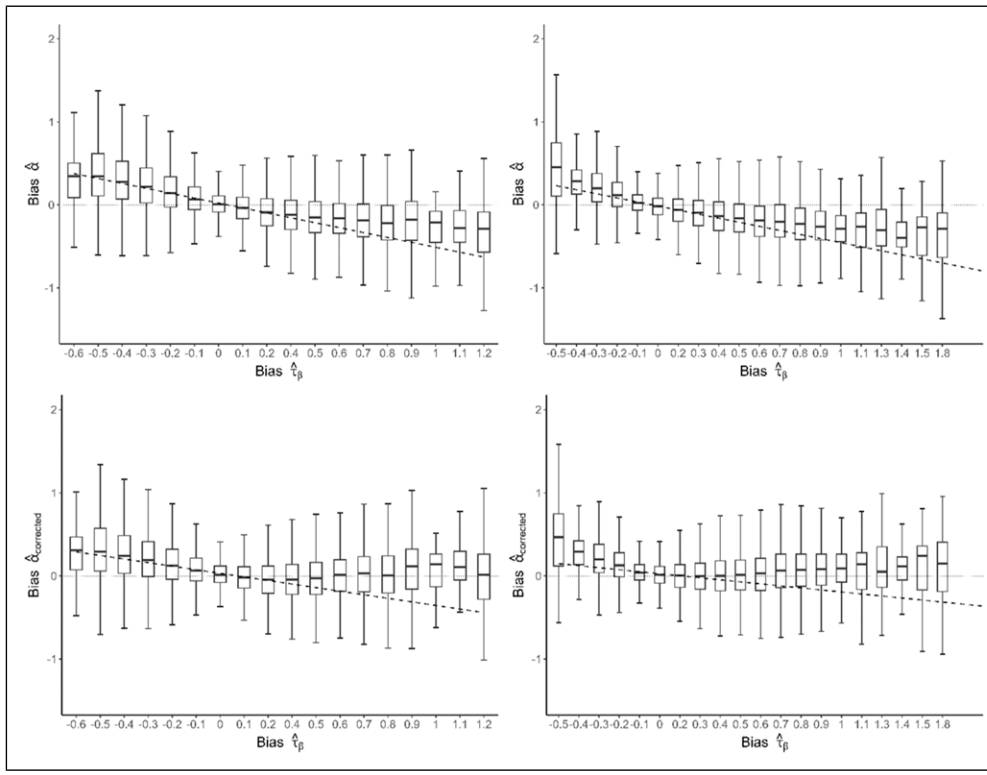


**Figure 7.** Bias in $\widehat{\alpha}_k$ Due to Bias in $\widehat{\tau}_\beta$ Before and After Applying the Correction Procedure.
Note. Top row: Bias in $\widehat{\alpha}_k$ due to bias in $\widehat{\tau}_\beta$. Bottom row: Bias in $\widehat{\alpha}_k$ due to bias in $\widehat{\tau}_\beta$ after application of the bias correction procedure. Left column: OH2PL. Right column: SHPL. Outliers not shown for legibility purposes. The dashed lines over the boxplots illustrate the regression of $\widehat{\alpha}_k$ (top row) and $\widehat{\alpha}_{k_{corrected}}$ (bottom row) on $\widehat{\tau}_\beta$.

discrimination parameter estimates independent from the true variance component of the difficulty parameters (an independent source of bias), which (2) increases the likelihood to benefit from the hierarchical approach even under unfavorable data conditions.

At first sight, a linear trend seemed to apply, but a closer look based on distribution-free measures as the series of boxplots (Figure 7, top row) revealed a certain non-linearity, that is, a logarithmic kind of association.

Because we knew the true values of all parameters in our simulation study, we could calculate the actual bias of both $\widehat{\alpha}_k$ and $\widehat{\tau}_\beta$. Exploring by means of nonlinear least squares estimation, how the bias of the individual parameters $\widehat{\alpha}_k$ depend on the bias of $\widehat{\tau}_\beta$, we identified the approximation

$$\tilde{b}_{\widehat{\alpha}} = -\frac{1}{2}\log(b_{\widehat{\tau_\beta}} + 1) \tag{15}$$

with $\tilde{b}_{\widehat{\alpha}}$ denoting the predicted bias. Supplement 2 shows that this correction worked exactly as expected. Moreover, Supplement 2 provides the details of how equation (15) has been derived.

However, in real-life applications the bias of $\widehat{\tau}_\beta$ remains unknown. Thus, we have to take a second step to find a suitable estimate for it in order to predict the bias of $\widehat{\alpha}_k$ (and correct for it, subsequently). In order to find a suitable approximation for the bias of $\widehat{\tau}_\beta$, namely, $b^*_{\widehat{\tau_\beta}}$, we set up a prediction model using information available in real-life applications

$$b^*_{\widehat{\tau_\beta}} = \beta_0 + \beta_1 N * \beta_2 K * \beta_3 M(\widehat{\tau}_\beta) * \beta_4 SD(\widehat{\tau}_\beta) * \beta_5 VAR(\widehat{\tau}_\beta) \tag{16}$$

where $N$ is the sample size, $K$ the number of items, $M(\widehat{\tau}_\beta)$ the mean, $SD(\widehat{\tau}_\beta)$ the standard deviation, and $VAR(\widehat{\tau}_\beta)$ the variance of the posterior distribution of the estimate of the variance component. The bottom row of Figure 7 (again, a series of boxplots summarizing the bias in the corrected $\widehat{\alpha}_k$ (the y-axis) at fixed cutpoints of the bias in $\widehat{\tau}_\beta$ (the x-axis) with a dashed line over the boxplots illustrating the regression of the corrected $\widehat{\alpha}_k$ on $\widehat{\tau}_\beta$) illustrates that the linearization of the bias of $\widehat{\alpha}_k$ was quite successful. We could reduce the bias in $\widehat{\alpha}_k$ for both the SH2PL ($\beta = -0.54, SE = 0.002$) and OH2PL ($\beta = -0.25, SE = 0.002$). Some non-linearity remains which was a consequence of the lack of approximation when $\widehat{\tau}_\beta$ is underestimated.

## Discussion and Conclusion

Our goal in this study was to provide an in-depth investigation of the question whether Bayesian hierarchical IRT models behave differently than their general counterparts in terms of the accuracy of the individual parameter estimates. We found (1) the Bayesian hierarchical specifications of the 2PL to yield individual parameter estimates consistently less biased compared to their nonhierarchical counterpart (especially in smaller samples), and (2) the bias in the individual item parameter estimates being partly independent from the bias in their associated true variance components. However, as shown by the relation between the bias in the individual discrimination parameter estimates and their true variance components, both are independent only when $\tau_\alpha \leq 0.4$. Considering that $\tau_\alpha$ is in many applications smaller than 0.4, our findings provide strong evidence that the performance of the Bayesian H2PL is in fact unique: The resulting item parameter estimates are not only more accurate (in terms of bias), but also more precise (in terms of HDI), compared to nonhierarchical approaches (for further justification regarding the increased precision, see also Jackman, 2009; Katahira, 2016). Our results also indicate, however, that this uniqueness is not a consequence of a generally different behavior, but rather a consequence of the interplay between item parameter and model characteristics.

Thus, from a theoretical point of view, the results of this study indicate that the connection between variance, shrinkage, and bias, a common characteristic of Bayesian hierarchical models (e.g., Rouder et al., 2017), albeit not completely absent, is not that pronounced in Bayesian hierarchical IRT models. In other words, shrinkage of the individual estimates towards their respective grand means does not lead, on average, to a marked increase in bias in the individual item parameter estimates. The difference between the results regarding the relative and absolute bias can be explained by the fact that only the latter explicitly captures the bias of discrimination parameters on the margins of the parameter distribution. Interestingly, even in terms of absolute bias the advantage of the hierarchical specifications over their nonhierarchical counterparts remains. Thus, while the behavior is in its core not different from general hierarchical models, the Bayesian hierarchical specifications of the 2PL provides a means to overcome the tradeoff between precision and accuracy of the individual item parameter estimates.

What does this rather theoretical finding mean for applied educational and psychological measurement? In the following, we briefly outline three consequences resulting from our finding that are relevant for applied IRT modeling.

First, using hierarchical Bayesian approaches for item calibration reduces item calibration error, one of the primary sources of biased ability estimates in computerized adaptive testing (CAT; e.g., Frey, 2023). More specifically, with the hierarchical Bayesian approach it is possible to avoid capitalization on chance in item selection due to spuriously large discrimination parameters (Patton et al., 2013). Given shrinkage, the overestimation of the item discrimination parameter is less likely to occur. As shown in this paper, the shrinkage associated with the item discrimination parameters does not lead to markedly biased parameter estimates in typical conditions (i.e., $\tau_\alpha \leq 0.4$). When $\tau_\alpha > 0.4$, however, the variance component is underestimated in both hierarchical specifications, which makes non-negligible bias more likely. This is also corroborated by the results regarding the absolute bias. Thus, $\tau_\alpha = 0.4$ can be considered as critical variance for item discrimination parameter estimates.

Second, consequently, using the hierarchical Bayesian approach is likely to avoid capitalization on item calibration error by the maximum information criterion in CAT (Patton et al., 2013). Since the item discrimination parameter plays a dominant role, unbiased parameter estimates are crucial for an accurate calculation of the Fisher information. Thus, the hierarchical Bayesian approach combined with the bias correction procedure outlined in this paper directly contributes to a more accurate calculation of the information contained in an item bank, especially in small samples. This translates into advantages regarding ability estimates and was shown by Wagner et al. (2022). Typically, item calibration error is largest when calibration samples are small; as shown in this paper, however, smaller sample sizes are not associated with larger calibration errors when utilizing the hierarchical Bayesian approach. This in turn leads to more flexibility when it comes to the calibration of new item banks with small samples, for example, when using continuous calibration methods (e.g., Fink et al., 2018).

Third, the benefits of using the hierarchical Bayesian approach are relatively independent of the specification of its prior structure. The advantages of the OH2PL over the standard Inverse Wishart specifications still exist, but they are small: both overestimate $\tau_\alpha$ when the variance of the item discrimination parameter estimates is smaller than .4 and underestimate $\tau_\alpha$ when the variance component is larger than .4. The underestimation is virtually indistinguishable across model specifications. This implies that, although the literature frequently discourages researchers from using Inverse Gamma or Inverse Wishart distributions (Gelman, 2006), the standard specification is a viable distribution in the IRT context. Moreover, the hierarchical Bayesian approach offers considerable flexibility when it comes to prior specification and structure (see also Koenig et al., 2022, for an investigation into the robustness of the performance of the OHPL under different prior specifications).

Our bias correction procedure is easy to apply. The prediction model for $b^{*}_{\widehat{\tau_\beta}}$ is included in the online supplementary material and can be used to predict $b^{*}_{\widehat{\tau_\beta}}$ for any test situation. The only things necessary are the data specifications and the posterior means and standard deviations of previously (or initially) calibrated item parameters. An illustration of how to apply the procedure is included in the online supplement. Moreover, the bias correction procedure can easily be extended to include other information, or can easily be integrated in similar efforts to reduce bias in item parameter estimates.

Taken together, the results of this study show that the curious behavior of the hierarchical Bayesian approach can be utilized to improve the accuracy and precision of the resulting item parameter estimates, not only in the context of the 2PL model, but also more complex models such as the GPCM. This in turn is beneficial for the precision of ability estimation and renders it especially appealing for situations where test information is crucial. Moreover, the hierarchical Bayesian approach facilitates applications of IRT models in situations that would not be feasible with alternative methods, for example, when recruiting large calibration samples is not possible (e.g., university exams or in clinical contexts).

To conclude, we could show that the characteristics of parameters typically found in applications of IRT models in combination with Bayesian hierarchical modeling indeed create a unique situation where the resulting item parameter estimates are not only more precise, but also more accurate, compared to nonhierarchical approaches. The contributions of our simulation study can serve as a reference for applied researchers on when and how to use Bayesian hierarchical approaches in IRT modeling without having to worry about potentially biased item parameter estimates. This should be appealing for a wide range of psychometric applications and psychological research.

## Acknowledgments

We would like to thank the Editor in Chief Dr John R. Donoghue and the anonymous reviewers for their valuable, constructive and helpful comments on our 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
Rainer W. Alexandrowicz https://orcid.org/0000-0001-9846-8936

## Supplemental Material

Supplemental material for this article is available online.

## References

Annis, J., Miller, B., & Palmeri, T. (2017). Bayesian inference with stan: A tutorial on adding custom distributions. *Behavior Research Methods*, *49*(3), 863–886. https://doi.org/10.3758/s13428-016-0746-9

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., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. In S. K. Updahyay, U. Singh, D. K. Dey, & A. Loganathan (Eds.), *Current trends in Bayesian methodology with applications* (pp. 79–102), Chapman & Hall.

Bezirhan, U., von Davier, M., & Grabovsky, I. (2021). Modeling item revisit behavior: The hierarchical speed-accuracy-revisits model. *Educational and Psychological Measurement*, *81*(2), 363–387. https://doi.org/10.1177/0013164420950556

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479), Addison-Wesley.

Carpenter, C., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & 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. (2023). *The theory and practice of item response theory* (2nd ed.), Guilford Press.

Efron, B., & Morris, C. N. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119–127. https://doi.org/10.1038/scientificamerican0577-119

Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, *60*, 327–346. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_3-2018_327-346.pdf

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

Frey, A. (2023). Computerized adaptive testing and multistage testing. In R. J. Tierney, F. Rizvi, & K. Erkican (Eds.), *International encyclopedia of education* (4th ed., pp. 209–216), Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10028-4

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., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2014). *Bayesian data analysis* (3rd ed.), CRC Press.

Gilholm, P., Mengersen, K., & Thompson, H. (2021). Bayesian hierarchical multidimensional Item Response Modeling of small sample, sparse data for personalized developmental surveillance. *Educational and Psychological Measurement*, *81*(5), 936–956. Advance online publication. https://doi.org/10.1177/0013164420987582

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

Hoffman, M., & Gelman, G. (2014). The No-U-Turn-Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(46), 1351–1381.

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

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*(1), 69–86. https://doi.org/10.1207/s15327906mbr2301_4

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, *73*(1), 37–58. https://doi.org/10.1016/j.jmp.2016.03.007

Koenig, 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

Koenig, C., Spoden, C., & Frey, A. (2022). Robustness of the performance of the optimized hierarchical two-parameter logistic IRT model for small-sample item calibration. *Behavior Research Methods*, *55*(8), 3965–3983. https://doi.org/10.3758/s13428-022-02000-5

Levy, R., & Mislevy, R. (2016). *Bayesian psychometric modeling*, Chapman & Hall.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–177. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Neal, R.M. (2011). MCMC Using Hamiltonian Dynamics. In S. Brooks, A. Gelman, G.L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 113–162). Chapman & Hall.

OECD. (2021). *PISA 2018 technical report*, OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/

Patton, J., Cheng, Y., Yuan, K.-H., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, *37*(1), 24–40. https://doi.org/10.1177/0146621612461727

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

Rouder, J. N., Morey, R. D., & Pratte, M. S. (2017). Bayesian hierarchical models of cognition. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung. *Cambridge handbooks in psychology. New handbook of mathematical psychology: Foundations and methodology* (Eds.), (pp. 504–551), Cambridge University Press. https://doi.org/10.1017/9781139245913.010

Stan Development Team. (2020). *Stan modeling language user's guide and reference manual*. Create Space Independent Publishing Platform, Version 2.19.1 [Computer software manual]. Retrieved from. https://mc-stan.org/

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 83–112. https://doi.org/10.1111/bmsp.12188

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Wagner, W., Zitzmann, S., & Hecht, M. (2022). *Multidimensional 1- and 2-parameter item response models in small (and large!) samples*. PsyArXiv. https://doi.org/10.31234/osf.io/tp6fy

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling*, *26*(4), 646–661. https://doi.org/10.1080/10705511.2018.1545232