AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000552

Bayesian Penalty Methods for Evaluating Measurement Invariance in Moderated Nonlinear Factor Analysis

Holger Brandt[1], Siyuan Marco Chen[2], and Daniel J. Bauer[2]
[1] Methods Center, University of Tübingen
[2] Department of Psychology, University of North Carolina at Chapel Hill

Abstract

Measurement invariance (MI) is one of the main psychometric requirements for analyses that focus on potentially heterogeneous populations. MI allows researchers to compare latent factor scores across persons from different subgroups, whereas if a measure is not invariant across all items and persons then such comparisons may be misleading. If full MI does not hold further testing may identify problematic items showing differential item functioning (DIF). Most methods developed to test DIF focused on simple scenarios often with comparisons across two groups. In practical applications, this is an oversimplification if many grouping variables (e.g., gender, race) or continuous covariates (e.g., age) exist that might influence the measurement properties of items; these variables are often correlated, making traditional tests that consider each variable separately less useful. Here, we propose the application of Bayesian Moderated Nonlinear Factor Analysis to overcome limitations of traditional approaches to detect DIF. We investigate how modern Bayesian shrinkage priors can be used to identify DIF items in situations with many groups and continuous covariates. We compare the performance of lasso-type, spike-and-slab, and global-local shrinkage priors (e.g., horseshoe) to standard normal and small variance priors. Results indicate that spike-and-slab and lasso priors outperform the other priors. Horseshoe priors provide slightly lower power compared to lasso and spike-and-slab priors. Small variance priors result in very low power to detect DIF with sample sizes below 800, and normal priors may produce severely inflated type I error rates. We illustrate the approach with data from the PISA 2018 study.

Translational Abstract

Measurement invariance (MI) is one of the main psychometric requirements for analyses that focus on populations with different subgroups. MI allows researchers to compare persons from these different subgroups, whereas if a measure is not invariant across all items and persons then such comparisons may be misleading. If full MI does not hold further testing may identify problematic items showing differential item functioning (DIF). Most methods developed to test DIF focused on simple scenarios often with comparisons across two groups. In practical applications, this is an oversimplification if many grouping variables (e.g., gender, race) or continuous covariates (e.g., age) exist that might influence the measurement properties of items. These variables are often correlated, making traditional tests that consider each variable separately less useful. Here, we propose the application of Bayesian Moderated Nonlinear Factor Analysis (MNLFA) to overcome limitations of traditional approaches to detect DIF. We investigate how modern Bayesian shrinkage priors can be used to identify DIF items in situations with many groups and continuous covariates. Results indicate that spike-and-slab and lasso priors outperform the other priors. Horseshoe priors provide slightly lower power compared to lasso and spike-and-slab priors. Small variance priors result in very low power to detect DIF with sample sizes below 800, and normal priors may produce severely inflated type I error rates.

Keywords: differential item functioning, shrinkage prior, nonlinear factor analysis, spike-and-slab, lasso

Measurement invariance (MI) is one of the most important psychometric properties for a test or questionnaire that is used in heterogeneous populations. If MI holds, results from tests can be compared, for example, across groups such as gender, or across persons with different scores in a continuous covariate such as age or socio-economic status (SES). If MI does not hold, persons with the same observed scores on a test might have different latent true scores; or persons with different observed test scores may actually be very similar with respect to their latent true scores. In both cases, comparisons of the test scores across persons are invalid because they are biased estimates of the true underlying scores persons have on the latent construct (Curran et al., 2018; Millsap, 2011).

MI is satisfied when the conditional distributions of the item responses are identical for all individuals with the same latent true scores. This requirement must be met simultaneously for all items for full invariance to hold. In practice, however, it is often the case that one or more items fail to show this property, resulting in partial

Holger Brandt https://orcid.org/0000-0001-5906-5202
Correspondence concerning this article should be addressed to Holger Brandt, Methods Center, University of Tübingen, Hauüerstr. 11, 72076 Tübingen, Germany. Email: holger.brandt@uni-tuebingen.de

invariance. These items are said to exhibit differential item functioning (DIF). They may be harder or easier to answer for some people, resulting in a location shift in the conditional distribution of the item responses that would manifest in different intercepts, or they may be more highly related to the latent factor for some people than others, manifested in different factor loadings. Fortunately, if DIF is correctly identified, it can be taken into account when calculating test scores, thus allowing for valid comparisons of test scores across persons of different subgroups even if MI does not completely hold (Byrne et al., 1989). For traditional DIF testing, one item needs to be used as anchor item. Anchor items are assumed to be DIF free (an assumption that may be violated, see below).

Many DIF detection procedures used in the recent past were based on the structural equation modeling (SEM) framework (cf. Bauer et al., 2020). In most cases, multiple sample analyses (MSAs) were conducted to contrast levels of a single grouping variable such as gender. However, this approach is limited to examination of one or maybe two grouping variables at a time. It is not amenable to the examination of many, potentially correlated grouping variables, which limits its practical use. For example, in large-scale data sets, heterogeneous populations are defined by many (correlated) categorical variables (e.g., countries, race, gender). Further, MSA does not easily admit continuous covariates (e.g., age or SES). Valid DIF detection methods need to be able to address such situations appropriately, that is, they need to take the complex pattern of correlated categorical and continuous variables into account. As an alternative method, MIMIC models ("multiple indicators multiple causes"; e.g., Barendse et al., 2010, 2012; Muthén, 1989; Oort, 1992) can be used to address DIF for multiple binary or continuous covariates; but, this method is limited to DIF in the intercepts and cannot be used for the detection of DIF in factor loadings (though see Woods & Grimm, 2011).

Recently, a moderated nonlinear latent factor analysis (MNLFA) was suggested that can be viewed as an extension of the MIMIC model (Bauer, 2017; Bauer & Hussong, 2009). The MNLFA allows researchers to detect DIF across several (correlated) grouping variables as well as across different levels of (several) continuous covariates. It can be used to test DIF in both intercepts and factor loadings in the presence of true differences in the latent mean and/or variance as a function of the covariates. Thus, this model extends possibilities to investigate DIF if partial MI holds.

While MNLFA has been conceptualized such that it can account for multiple variables simultaneously, the number of parameters that need to be tested grows rapidly because it depends both on the number of items that are tested for DIF, and the number of covariates that cause DIF. Traditional approaches to test complex models in such data situations have severe limitations. In principle, they either use a sequence of tests for the parameters—which implies similar problems as step-wise regression methods (e.g., Bauer et al., 2020; Derksen & Keselman, 1992). Or, if all parameters are tested simultaneously, they might break down or at least provide results with low accuracy and power (e.g., in nonlinear SEM, see Brandt et al., 2018). In both situations, a reliable assessment of DIF is hindered and alternatives are needed.

Scope

In this article, we will explore the possibilities of a Bayesian extension of the MNLFA to account for large numbers of groups and covariates. This will extend previous work on Bayesian MNLFA that focused on few groups in an initial exploratory study (Chen et al., 2021). We will introduce different Bayesian shrinkage priors that can be used for model sparsity in these situations. We will discuss different modern shrinkage priors, namely the global–local priors and spike-and-slab priors (e.g., Carvalho et al., 2009; Ishwaran & Rao, 2005), and compare them to traditional shrinkage methods such as the (adaptive) Bayesian lasso (Park & Casella, 2008; Tibshirani, 1996). In an extensive simulation study, we will investigate how such priors perform in data sets with many groups and covariates as it is more common in, for example, large-scale data sets. We will show how these modern shrinkage priors outperform standard Bayesian implementations with normal priors and which of them is best suited for DIF detection.

Outline

In the next section, we will discuss different traditional and modern shrinkage priors; we will provide advantages and limitations of the methods. Then, we will introduce the Bayesian MNLFA implementation. We will present a simulation study that compares the performance of standard priors (e.g., normal) to modern shrinkage priors (Bayesian lasso, horseshoe or Dirichlet Laplace priors, and spike-and-slab priors) in Bayesian MNLFA. We illustrate the Bayesian MNFLA with an empirical data set within an educational psychology context. We will end with a discussion of the results and their ramifications.

Penalized Estimation With Shrinkage Priors

Penalization methods were developed for situations in which, for example, many predictors are used to predict an outcome, or for semi-parametric models (e.g., smoothing splines) where the number of parameters depends on the amount of smoothing that is allowed (e.g., Guo et al., 2012; Tibshirani, 1996; Wood, 2003). One of the original models developed for such models in the regression framework is the lasso ("least absolute shrinkage and selection operator") that is an extension of the ordinary least squares (OLS) estimator (Tibshirani, 1996). For a centered dependent variable $\tilde{\mathbf{y}}$ and a set of q predictor variables $\mathbf{X}$, the lasso obtains estimates via minimizing:

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \left\{ (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \phi \sum_{k=1}^{q} |\beta_k| \right\}, \tag{1}$$

with a shrinkage factor $\phi \geq 0$ that controls the amount of shrinkage, and regression coefficients β;. For $\phi = 0$, no shrinkage is imposed and the original OLS estimator results. The larger ϕ is, the more shrinkage is imposed on β.

One of the main advantages of the lasso is that it shrinks small effects exactly to zero and thus can be used as a selection tool for predictors. One of the main disadvantages of the lasso is that it can induce bias in coefficients that are nonzero in the population because the same shrinkage factor is used for all coefficients (Zou, 2006). The adaptive lasso was developed to overcome this problem by using different shrinkage factors ϕ_k for each of the $k = 1 \ldots q$ regression coefficients (Zou, 2006). Other similar shrinkage methods that tried to overcome this problem include the "smoothly clipped absolute deviation penalty" (SCAD) and the "minimax

concave penalty" (MCP) that both use specific spline functions for the penalty term (Fan & Li, 2001; C. H. Zhang, 2010).

For both lasso and adaptive lasso (as well as SCAD and MCP) in their frequentist implementation, one of the main challenges is the actual choice of the shrinkage factors that can be determined with different methods such as shrinkage plots, cross-validation, or information criteria like the Bayesian Information Criterion (BIC) (e.g., Jacobucci et al., 2016; Lykou & Ntzoufras, 2013; Tutz & Schauberger, 2015). For each of these methods, several models with different shrinkage factors need to be estimated and compared. Then, the model is selected, for example, that has the smallest BIC. For more complex models, this approach is computationally burdensome even if only the standard lasso is used with a single shrinkage factor to determine. For the adaptive lasso, it seems not feasible to estimate several models for each of the several shrinkage factors for each coefficient (e.g., 30 ϕs for 100 predictors).

Bayesian Shrinkage Priors

In the Bayesian framework, a variety of different shrinkage priors were developed. Traditional shrinkage priors are Bayesian equivalents of the frequentist lasso. From a conceptual perspective, their shrinkage properties are less beneficial compared to more modern shrinkage priors as they have been developed in the last decade. These modern priors have a more complicated setup that use different types of mixture distributions (e.g., Bhadra et al., 2017; Ishwaran & Rao, 2005; Polson & Scott, 2010). They can be categorized into one- and two-component models. In the following, we will discuss the conceptual setup of these three types of shrinkage priors.

Traditional Shrinkage Priors

The Laplace prior is the first Bayesian implementation of a shrinkage prior and it is the direct adaptation of the (adaptive) lasso (Hans, 2009; Leng et al., 2014; Park & Casella, 2008). It uses a double exponential or Laplace prior where its mode of the posterior distribution is equivalent to the lasso. The Laplace prior for a coefficient θ_j can be expressed as (e.g., Leng et al., 2014):

$$\theta_j \sim \text{dexp}(\sigma/\phi_{(j)}) \quad \phi^2_{(j)} \sim \text{Ga}(a_1, b_1) \quad \sigma^{-2} \sim \text{Ga}(a_2, b_2), \tag{2}$$

where dexp(s) is the double exponential distribution with zero mean and scale s. σ is a scaling parameter and $\phi_{(j)}$ is the shrinkage parameter. The prior for the (square of the) shrinkage factor $\phi^2_{(j)}$ is a Gamma distribution, Ga(a, b), with hyperparameters a, b (Hans, 2009, 2010; Park & Casella, 2008). For small values, no shrinkage is imposed (wide double exponential distribution); for larger values, the shrinkage increases (sharper double exponential distribution). In its original implementation, a single value $\phi_{(j)} = \phi$ for all coefficients was used (Hans, 2009; Park & Casella, 2008). Later, the implementation was extended to the adaptive lasso version where each coefficient has a separately sampled shrinkage factor ϕ_j (Leng et al., 2014).

One of the main benefits of the Bayesian version of the lasso is that the shrinkage factor can be derived from the data by using prior distributions for $\phi_{(j)}$. Often, the mean of the posterior (instead of the mode) is used to assess the parameters. In this case, it cannot be expected that small coefficients are shrunken exactly to zero but only close to zero.

Two-Component Models

The two-component model is often called spike-and-slab prior (George & Culloch, 1993; Ishwaran & Rao, 2005; Mitchell & Beauchamp, 1988; Piironen & Vehtari, 2017). It is conceptualized as a mixture distribution that has one mixture component that is a very dense mass or often even a point mass around zero (spike) and a second part that is a wider distribution for those coefficients that are non-zero (slab). For a parameter θ_j, the prior can be described as (see Bhadra et al., 2017; Ishwaran & Rao, 2005):

$$\theta_j \sim \underbrace{(1-\phi_j)N(0, \sigma_1^2)}_{\text{spike}} + \underbrace{\phi_j N(0, \sigma_2^2)}_{\text{slab}}, \tag{3}$$

where $N(0, s^2)$ is the normal distribution with zero mean and variance s^2. $\sigma_1^2 \ll \sigma_2^2$ are scaling parameters for the spike and slab parts; ϕ is either a 0–1 variable (Bernoulli distributed) or a Beta distributed variable which can be interpreted as weight or probability (e.g., Bhattacharya et al., 2015; Hans, 2010; Ishwaran & Rao, 2005). If ϕ_j is Bernoulli distributed it can be interpreted as an indicator for whether a specific predictor is zero or not. If it is Beta distributed, it can be interpreted as complexity parameter, that is, the probability that a coefficient is not zero ($\phi_j = P(\theta_j \neq 0)$; Ishwaran & Rao, 2005). Although different to the shrinkage factor in the Bayesian lasso above, ϕ_j has the same role in spike-and-slab priors, that is to shrink and select coefficients.

With regard to the scaling parameter of the spike part, σ_1^2 is often chosen to be zero, which results in a delta spike at zero (i.e., an actual point mass, δ_0), so that the resulting prior is a mixture of a discrete value and a normal distribution. Alternatively, small values could be chosen for σ_1^2 so that the mixture is composed of two normal distributions (for a discussion of its benefits see Ishwaran & Rao, 2005). For σ_2^2 (the scaling parameter of the slab), either a fixed value can be used, or an additional prior such as an exponential prior is included, which then results in a double exponential distribution for the slab (e.g., Hans, 2010).

From a conceptual perspective, the advantage of the spike-and-slab prior over the lasso type priors is that it has better shrinkage properties because it can actually shrink coefficients exactly to zero via the spike part. The spike-and-slab prior is sometimes cited as the gold standard of Bayesian shrinkage priors (Bhadra et al., 2017; Piironen & Vehtari, 2017). A disadvantage of the prior is that the resulting parameter space is huge if ϕ_j is sampled from a Bernoulli distribution and many parameters are part of the model, which can result in difficulties in sampling and convergence.

One-Component Models

One-component models or global–local priors are conceptualized in the following way (Bhattacharya et al., 2015; Polson & Scott, 2010):

$$\theta_j \sim N(0, \phi^2\tau_j^2) \quad \tau_j \sim f \quad \phi \sim g, \tag{4}$$

where f and g are specific distributions that define the priors within this class. All global–local priors have in common that they have a global shrinkage part that controls the overall shrinkage (ϕ), that is, how sparse the complete parameter vector is (similar to the shrinkage factor in the Bayesian lasso above). And, they have a

second local shrinkage part (τ_j) that allows individual parameters to escape the shrinkage (e.g., Bhadra et al., 2017; Polson & Scott, 2010). This setup keeps larger parameters theoretically unbiased (because they are not shrunken towards zero as in the lasso type priors). Mainly two different priors have received attention in recent years: the horseshoe (and its extensions) and the Dirichlet Laplace prior.

The horseshoe prior (Carvalho et al., 2009, 2010) is formulated as:

$$\theta_j \sim N(0, \phi^2\tau_j^2) \quad \tau_j \sim C(0, 1)^+ \quad \phi \sim C(0, a)^+, \tag{5}$$

where $C(0, s)^+$ is the Half-Cauchy distribution with zero mean and scale s. The heavy-tailed Half-Cauchy prior for τ_j ensures that sufficiently large numbers can be sampled for coefficients that are actually nonzero.

Two extensions of the horseshoe were proposed: the regularized horseshoe (Piironen & Vehtari, 2017) and the horseshoe+ (Bhadra et al., 2017). The regularized horseshoe adds a regularization to large coefficients, which in some situations has advantages (e.g., in logistic regressions where weak identification issues can occur). The horseshoe+ prior adds an additional level in the prior hierarchy that theoretically improves the shrinkage properties of the prior.

The horseshoe prior has been investigated primarily for regression models with many predictors and it has been shown to perform very well in these scenarios. Theoretically, the horseshoe prior has better shrinkage characteristics than the Bayesian lasso (Bhadra et al., 2017). The horseshoe prior is very similar to the spike-and-slab prior if a delta spike is used (i.e., $\sigma_1^2 = 0$ in Equation 3). In this case, the key difference between the two priors is that τ_j has a Half-Cauchy distribution in the horseshoe, while it has a Beta or Bernoulli distribution in spike-and-slab priors (Piironen & Vehtari, 2017).

Another global–local prior is the Dirichlet Laplace (DL) prior (Bhattacharya et al., 2015; Y. Zhang & Bondell, 2018) that uses a Dirichlet distribution for the local shrinkage:

$$\begin{aligned} \theta_j &\sim \text{dexp}(\tau_j\phi) \quad \tau_1, \ldots, \tau_p \sim \text{Dir}(a_\tau, \ldots, a_\tau) \\ \phi &\sim \text{Ga}(pa_\tau, 1/2), \end{aligned} \tag{6}$$

where $\text{Dir}(a_\tau, \ldots, a_\tau)$ is the Dirichlet distribution with hyperprior a_τ (a multivariate extension of the Beta distribution) and p is the number of parameters.[1] τ_j is the local shrinkage part that indicates whether a parameter is zero or not. Small values of the hyperprior a will result in sparse parameter vectors (with many zeros). ϕ again controls the global shrinkage and is sampled from a Gamma distribution.

Several other priors were developed such as normal exponential-gamma (Griffin & Brown, 2010), normal-gamma gamma, or double Pareto priors (Armagan et al., 2013). The generalized Beta-mixtures (Armagan et al., 2011) or normal-gamma gamma priors (Griffin & Brown, 2010) are more general frameworks that comprise as special cases horseshoe or normal-exponential-gamma priors.

Modern Priors in Latent Variable Models

So far, only few implementations of modern shrinkage methods exist for latent variable models. Most implementations used frequentist lassos (Bauer et al., 2020; Belzak & Bauer, 2020; Huang, 2020; Jacobucci et al., 2016) that have limitations such as inconsistent estimates due to the single shrinkage factor for all coefficients (Zou, 2006) and that many models need to be estimated in order to derive the shrinkage factor. Bayesian implementations have focused on adaptive lasso implementations via the Laplace prior (Feng et al., 2015; Guo et al., 2012). Brandt et al. (2018) tested a spike-and-slab prior that used a double exponential distribution for the slab, and a Beta-distributed complexity parameter (ϕ_j) that reduced the computational burden compared to the Bernoulli-distributed prior discussed above. For latent variable models with interaction effects, they found that this implementation outperformed standard frequentist approaches and the adaptive Bayesian lasso. A similar result was obtained in a two-parameter logistic item response theory model where the authors showed that spike-and-slab priors were more reliable for DIF detection than a frequentist lasso implementation in a situation with three background variables (Chen et al., 2021).

A different type of prior in latent variable models that can be subsumed under the classification of shrinkage type priors was suggested by Shi et al. (2017). They used a normal prior with small variance such as $N(0, 0.001)$ to aid in selecting anchor items in MI testing. This prior can be seen as spike-and-slab prior without the slab. From a conceptual perspective, this prior should produce problematic (biased) estimates for coefficients that are non-zero in the population. As a consequence, the authors suggested to re-estimate the model using standard priors after having decided upon the anchor item. If, however parameter estimates are severely biased (i.e., parameter estimates are all shrunk to zero), the selection of anchor items might be biased, too, because DIF might be overlooked. In line with these expectations, Kim et al. (2017) showed in a larger study on MI that small variance priors are problematic in situations with larger DIF.

Modern shrinkage priors such as the horseshoe or DL priors have rarely been used in latent variable models so far. There are at least some indications that the results obtained for these priors in regression type models cannot be directly transferred to latent variable models. For example, multicollinearity might be a more severe issue in MNLFA than it is in regression models.

Bayesian MNLFA

MNLFA can be viewed as a generalization of both traditional MSA and MIMIC models. In contrast to MSA, it allows one to include both correlated grouping variables and continuous covariates. In contrast to MIMIC models, all parameters (including factor loadings and variances) can be modeled as a function of the covariates. Here, we first present the model equations and discuss the challenges that come with this modeling approach. Subsequently, we illustrate its Bayesian implementation.

Model

The general MNLFA measurement model is given by

$$\begin{aligned} g(\boldsymbol{\mu}_i) &= \boldsymbol{\nu}_i + \boldsymbol{\Lambda}_i\boldsymbol{\eta}_i, \\ (\mathbf{y}_i|\boldsymbol{\eta}_i) &\sim F(\boldsymbol{\mu}_i, (\boldsymbol{\Sigma}_i)), \end{aligned} \tag{7}$$

[1] The equation is a special case of Equation 4 because a double exponential distribution dexp(s) is equivalent to a normal distribution $N(0, s^2\upsilon)$, $\upsilon \sim \text{exp}(1/2)$, where $\text{exp}(\cdot)$ is the exponential distribution.

where $i = 1, \ldots, N$ indicates the person, $\mathbf{y}_i$ is a $p \times 1$ vector of the observed item scores. $(\mathbf{y}_i|\boldsymbol{\eta}_i)$ is the conditional distribution of $\mathbf{y}_i$ with conditional mean vector $\boldsymbol{\mu}_i = E[\mathbf{y}_i|\boldsymbol{\eta}_i, \mathbf{x}_i]$ and conditional covariance matrix $\boldsymbol{\Sigma}_i = V(\mathbf{y}_i|\boldsymbol{\eta}_i, \mathbf{x}_i)$. ν_i and $\boldsymbol{\Lambda}_i$ are person-specific intercept and factor loading matrices of dimensions $p \times 1$ and $p \times r$, respectively. $\boldsymbol{\eta}_i$ is an $r \times 1$ vector of latent factor scores for person i. g and F are a link and a distribution function, respectively, that depend on the type of data. For example, if $\mathbf{y}_i$ is continuous, g is the identity function and F is the multivariate normal distribution. If $\mathbf{y}_i$ is binary, g is the logit function and F is the multivariate Bernoulli distribution. Depending on the distribution function, the specification of individual parameters in $\boldsymbol{\Sigma}_i$ is either necessary or not (as indicated with the brackets in Equation 7).

In this general formulation, the latent factors have a distribution of $\boldsymbol{\eta}_i \sim MVN(\boldsymbol{\alpha}_i, \boldsymbol{\Psi}_i)$, where both the factors' mean vector and covariance matrix are person-specific.

DIF Detection

In order to detect DIF across a set of q continuous covariates or dummy-coded (categorical) grouping variables $\mathbf{x}$ ("coding variables"), the following model is specified for the subject-specific intercept vector:

$$\boldsymbol{\nu}_i = \boldsymbol{\nu}_0 + \mathbf{K}\mathbf{x}_i, \tag{8}$$

where ν_0 is a $p \times 1$ vector of baseline intercept values that hold when all continuous covariates and coding variables take on values of zero. $\mathbf{K}$ is a $p \times q$ matrix and its elements κ_{kj} indicate whether the kth variable y_k has DIF due to covariate x_j with respect to the intercept parameter κ.

For the factor loading matrix $\boldsymbol{\Lambda}_i$, the following model is used. For $a = 1, \ldots, r$, define $\boldsymbol{\lambda}_{ai}$ as the ath column of $\boldsymbol{\Lambda}_i$ that refers to the ath latent factor η_a. Then the factor loadings can be represented as

$$\boldsymbol{\lambda}_{ai} = \boldsymbol{\lambda}_{a0} + \boldsymbol{\Omega}_a\mathbf{x}_i, \tag{9}$$

where $\boldsymbol{\lambda}_{a0}$ is a $p \times 1$ vector of baseline factor loadings on the ath factor that hold when all continuous covariates and coding variables are zero. $\boldsymbol{\Omega}_a$ is the $p \times q$ matrix. Its elements ω_{akj} indicate the difference between the reference factor loading λ_{a0kj} and the kth factor loading y_k on η_m for covariate x_j. Any such difference represents loading DIF. Note that the factor loadings in $\boldsymbol{\Omega}_a$ are conceptually interaction effects as the factor loadings relate to the product of latent factors $\boldsymbol{\eta}$ and covariates $\mathbf{x}$.

In traditional terminology, metric (or weak) invariance holds if all coefficients in $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_r$ are equal to zero. Similarly, scalar (or strong) invariance holds if in addition all coefficients in $\mathbf{K}$ are zero. DIF for an item is indicated by any nonzero value in these coefficients.

Structural Invariance

Structural invariance testing can be used to investigate how latent factor means and variances differ across groups or how they depend on other covariates. In order to allow for differences in the latent means and variances, the following model is used (Bauer, 2017) for each of the $a = 1, \ldots, r$ factors:

$$\alpha_{ai} = \alpha_{a0} + \boldsymbol{\Gamma}_a\mathbf{x}_i, \tag{10}$$

$$\psi_{(aa)i} = \psi_{(aa)0} \exp(\beta_{(aa)}\mathbf{x}_i), \tag{11}$$

where $\boldsymbol{\Gamma}_a$ and $\beta_{(aa)}$ are $q \times 1$ vectors that indicate invariance if they are zero. Any nonzero elements indicate actual differences in the distribution of the latent variable as a function of the covariates/coding variables. Such differences, sometimes are referred to as "impact," do not represent a violation of MI and may be of substantive interest. They are only valid to be interpreted if all DIF that is present in the data has been correctly identified and included in the model. Note that $\psi_{(aa)\,i}$ and $\psi_{(aa)\,0}$ are the diagonal elements of the covariance matrix $\boldsymbol{\Psi}_i$ and $\boldsymbol{\Psi}_0$, respectively. Person-specific covariances between two factors a and a' ($\psi_{(aa')\,i}$) can be modeled similarly using a Fisher-z-transformation (see details in Bauer, 2017).

Considerations for the Bayesian Implementation of the MNLFA

One of the main challenges to detect DIF with the MNLFA is model complexity. The number of potential DIF parameters increases both with the number of items and the number of covariates. For example, in a situation with $p = 6$ indicators for a single factor and $q = 4$ covariates, a saturated model includes $p = 6$ intercepts in ν_0 and $p = 6$ factor loadings in $\boldsymbol{\lambda}_{10}$ (see Equations 8 and 9). Additionally, there are $p \times q = 24$ potential DIF parameters for the intercepts, $1 \times p \times q = 24$ potential DIF parameters for the factor loadings, and $2 \times q = 8$ potential impact parameters in $\boldsymbol{\Gamma}_1$, $\boldsymbol{\beta}_{(11)}$ (a total of 68 parameters, assuming all DIF effects are linear). Attempting to estimate all or even most of these parameters simultaneously using maximum likelihood can result in estimation problems and a lack of sensitivity for detecting DIF effects. In many situations, most of these DIF parameters (except of ν_0, $\boldsymbol{\lambda}_{10}$) will be zero.

A second main challenge is related to anchor items. In traditional methods for DIF detection, constraints need to be introduced into the model for identification. An anchor item needs to be chosen for each factor for which all elements in $\mathbf{K}$ and $\boldsymbol{\Omega}$ are set to zero, implying that this item does not have DIF. Several methods have been developed on how to select these anchor items (for a discussion see e.g., Belzak & Bauer, 2020). In general, there are severe limitations with these methods, and DIF detection for the remaining items is always conditional on the correct choice of the anchor item. Equivalently, structural invariance could be assumed[2] instead of an anchor item, but this simply shifts the problem and results in an assumption about latent factor means and variances that might be implausible and/or violated. In a Bayesian framework, Shi et al. (2017) suggested to use small variance priors for the factor loadings and intercepts which provide approximate identification in order to avoid the choice of anchor items. To achieve approximate identification, the priors concentrate the posterior toward a sparsity of DIF, with one or more items functioning as (approximate) anchors. In some cases, this may provide an unsatisfying solution to the anchor item selection problem. For instance, if there is only one anchor item, any item could be chosen and this would generate an equivalent likelihood. The priors could constrain the model toward selecting a particular item as the anchor, but this would remains a somewhat arbitrary choice. Given this arbitrariness, when only one or very few items function as anchors, typically the conclusion is that there is a failure of MI and that scores are

[2] By setting the coefficients in $\boldsymbol{\Gamma}_a$ and $\boldsymbol{\beta}_{(aa)}$ to zero and one, respectively.

incomparable. However, when sparsity of DIF is achieved and multiple anchors are identified (per covariate), then one may have greater confidence in the validity of the partially invariant solution and the comparability of scores adjusted for DIF (for a very general discussion of this topic, see San Martin & Gonzalez, 2020). This idea has recently been adapted for frequentist shrinkage estimation in DIF settings (Bauer et al., 2020).

We believe Bayesian MNLFA holds promise for helping to resolve these two challenges, but there are several aspects of its implementation that need attention to ensure optimal sensitivity to detect DIF. First, different types of parameters (e.g., factor loadings and intercepts) are tested for DIF and these parameters are not necessarily on equivalent scales. For example, DIF in intercepts is operationalized via main effects of the covariates ($\mathbf{K}\mathbf{x}_i$ in Equation 8), while DIF in factor loadings is operationalized via interactions between latent factors and covariates ($\mathbf{\Omega}_a\mathbf{x}_i\eta_{ai}$ in Equations 7 and 9). The actual size of the coefficients is relative to the size of variance of these different terms (Bohrnstedt & Goldberger, 1969) that may be very different even if covariates are standardized because the variance of product terms depends on higher moments of variables' distribution. Thus, a reliable method needs to be able to be able to identify relevant effect sizes even if the regression coefficients are on different scales (e.g., the adaptive Bayesian lasso). The performance of a non-adaptive Bayesian lasso might depend on the actual number of items with DIF and their pattern (i.e., the sparsity of the parameter vector). Initial results from a frequentist lasso version of the MNLFA supported this claim (Bauer et al., 2020). Second, the method should have similar selection properties as the original lasso that allows researchers to make unambiguous decisions about DIF. In the context of MNLFA, a huge number of parameters needs to be evaluated with the goal to identify a sparse structure (i.e., in optimal situations, many zeros in the matrices $\mathbf{K}$ and Ω_a). Especially modern shrinkage priors such as the spike-and-slab prior can account for this aspect. Third, while obtaining sparsity is very important, the actual power to detect DIF should be sufficient under empirically relevant scenarios (i.e., with many covariates). So far, some shrinkage priors were investigated for a two-parameter logistic-model with few background variables (Chen et al., 2021). In the context of many covariates and/or groups that results in far more complex models, it still remains unclear which of the shrinkage priors presented above has the highest relative power. Lastly, the method should avoid the problems associated with the selection of anchor items. That is, ideally, the method should not require anchor items to be selected a priori, but rather allow these to be detected in the course of estimating the model. This will be achieved by using shrinkage priors for the relevant parameters, which will result in approximate identification similar to the method proposed by Shi et al. (2017) with small variance priors. Here, we will present a general implementation of MNLFA without the use of a-priori-designated anchor items, and then shortly describe the different shrinkage priors that will be implemented. In the following section, these shrinkage priors will be tested.

Bayesian Model Implementation

It is straightforward to implement the MNLFA in a Bayesian framework. The observed variables' distributions can be specified as

$$y_{ik} \sim N(\mu_{ik}, \sigma^2_{ik}), \quad i = 1, \ldots, N, k = 1, \ldots, p \tag{12}$$

for continuous data, or

$$y_{ik} \sim \text{Bern}\big(\text{expit}(\mu_{ik})\big), \quad i = 1, \ldots, N, k = 1, \ldots, p \tag{13}$$

for binary data with $\text{expit}(x) = \exp(x)/(1 + \exp(x))$ (see Equation 7). The multivariate distribution of the latent factors is given by

$$\boldsymbol{\eta}_i \sim MVN(\boldsymbol{\alpha}_i, \boldsymbol{\Psi}_i), \quad i = 1, \ldots, N \tag{14}$$

The priors for the parameters that hold across all persons (indicated with zero subscripts in Equations 8–11) are given by

$$\nu_{0k} \sim N(\mu_{\nu 0k}, \sigma^2_{\nu 0k}), \quad k = 1, \ldots, p \tag{15}$$

$$\lambda_{a0k} \sim N(\mu_{\lambda a0k}, \sigma^2_{\lambda a0k}), \quad k = 1, \ldots, p, \ a = 1, \ldots, r \tag{16}$$

$$\alpha_{a0} \sim N(\mu_{\alpha a0}, \sigma^2_{\alpha a0}), \quad a = 1, \ldots, r \tag{17}$$

$$\boldsymbol{\Psi}_0^{-1} \sim \text{Wish}(\boldsymbol{\Psi}_0^*, df_{\Psi 0}), \tag{18}$$

where Wish is the Wishart distribution for precision matrix $\boldsymbol{\Psi}_0^{-1}$ (as it used in Jags). In the case of continuous items, the prior for the residual variance is given by

$$\sigma_k^{-2} \sim \text{Ga}(a_{\sigma k}, b_{\sigma k}), \quad k = 1, \ldots, p. \tag{19}$$

where σ_k^{-2} is the precision (again as it used in Jags).

Priors for the remaining coefficients $\mathbf{K}$, $\mathbf{\Omega}_a$, $\mathbf{\Gamma}_a$, and $\boldsymbol{\beta}_{(aa)}$ will be implemented via shrinkage priors. As there are many different shrinkage priors and only little information available about their performance in complex latent variable models, we decided to implement methods from all groups of shrinkage priors. We used standard normal priors as a baseline, small variance normal priors (Shi et al., 2017), traditional lasso type shrinkage priors (adaptive and nonadaptive Laplace priors; Hans, 2009; Leng et al., 2014; Park & Casella, 2008), one-component global-local type priors (horseshoe and DL priors; Bhattacharya et al., 2015; Carvalho et al., 2009), and two-component spike-and-slab type priors (Brandt et al., 2018; Ishwaran & Rao, 2005). A summary of the priors is presented in Table 1. More technical details on their implementation can be found in "Appendix A."

Simulation Study

In this section, we investigate how the different types of priors perform for the Bayesian MNLFA. We compare the performance of normal priors to different shrinkage priors, including small variance priors, different types of lasso type priors, global-local priors, and spike-and-slab priors.

Population Model

Data were generated based on a measurement model for a single factor model with $p = 6$ items $\mathbf{y}$ and q covariates $\mathbf{x}$. All covariates were correlated among each other and with the latent factor η with a correlation ρ.

Data for the items were generated by

$$\mathbf{y}_i = \boldsymbol{\nu}_i + \boldsymbol{\lambda}_{1i}\boldsymbol{\eta}_i + \boldsymbol{\delta}_i, \tag{20}$$

Table 1
Prior Specifications

Prior	Hierarchy			
$N(0, 1)$	$\theta_{kj} \sim N(0, 1)$			
$N(0, 0.001)$[a]	$\theta_{kj} \sim N(0, 0.001)$			
Blasso[b]	$\theta_{kj} \sim \text{dexp}(\sigma_k\phi)$	$\phi^{-2} \sim \text{Ga}(a, b)$		
aBlasso[c]	$\theta_{kj} \sim \text{dexp}(\sigma_k\phi_{kj})$	$\phi_{kj}^{-2} \sim \text{Ga}(a, b)$		
Horse[d]	$\theta_{kj} \sim N(0, \tau_{kj}^2\phi_k^2)$	$\tau_{kj} \sim C^+(0, c)$	$\phi_k \sim C^+(0, \text{c})$	
DL[e]	$\theta_{kj} \sim \text{dexp}(\tau_{kj}\phi_k)$	$\tau_{kj} \sim \text{Dir}(1\ldots1)$	$\phi_k^{-2} \sim \text{Ga}(q, 0.5)$	
N-spike[f]	$\theta_{kj} \sim N(0, \tau_{kj}\nu_{kj}^2)$	$\tau_{kj} \sim (1-\phi_{kj})\delta_{0.005} + \phi_{kj}\delta_1$	$\phi_{kj} \sim \text{Beta}(d, e)$	$\nu_{kj}^{-2} \sim \text{Ga}(a, b)$
δ-spike[g]	$\theta_{kj} \sim (1-\phi_{kj})\delta_0 + \phi_{kj}\text{dexp}(\sigma_k\tau_{kj})$	$\tau_{kj}^{-2} \sim \text{Ga}(a, b)$	$\phi_{kj} \sim \text{Beta}(d, e)$	

Note. θ_{kj} is the element in $\mathbf{K}$ or $\mathbf{\Omega_1}$, respectively, that represents DIF for the kth indicator variable ($k = 1, \ldots, p$) due to the jth covariate ($j = 1, \ldots, q$). a, b, c, d, and e are the hyperparameters. $N(0, 1)$ = normal prior; $N(0, 0.001)$ = small variance prior; Blasso = Bayesian Lasso; aBlasso = adaptive Baysian Lasso; Horse = horseshoe prior; DL = Dirchlet Laplace; N-spike = continuous spike-and-slab prior; δ-spike = categorical spike-and-slab prior; DIF = differential item functioning.
[a]Shi et al. (2017), [b] Hans (2009), [c] Leng et al. (2014), [d] Carvalho et al. (2010), [e] Bhattacharya et al. (2015), [f] Ishwaran and Rao (2005), [g] Brandt et al. (2018).

with $\eta \sim N(0, 1)$ and residuals $\boldsymbol{\delta} \sim N(\mathbf{0}, \mathbf{I}_p)$. Intercept and factor loading matrices were specified according to Equations 8 and 9 with $\nu_0 = \mathbf{0}_p$ and $\boldsymbol{\lambda}_{10} = \mathbf{1}_p$ as the population level p-dimensional vectors for intercept and factor loadings at $\mathbf{x}_i = \mathbf{0}_q$. Items without DIF thus had a reliability of 0.5, which is a realistic size for applied researchers.

Simulation Conditions

Sample Size

Sample sizes were set to $N = 200$, 400, and 800 (fixed design factor A). We decided to exclude larger sample sizes because the importance of priors compared to the likelihood component for the posterior diminishes and all priors theoretically may lead to very similar estimation results for large sample sizes (e.g., above 1,000).

Type of DIF

DIF was induced for intercepts only ($\mathbf{K}$) or simultaneously for both intercepts and factor loadings ($\mathbf{K}$ and $\mathbf{\Omega}_1$; fixed design factor B). We did not include a situation where only factor loadings produced DIF as this is an uncommon scenario.

Size of DIF

The size of the DIF was set to a standardized effect size of 0.1 versus 0.2 (small vs. medium; fixed design factor C) for the respective nonzero elements in $\mathbf{K}$ and/or $\mathbf{\Omega}_1$. These coefficients were in line with typical standardized effect size (cf. Bauer et al., 2020).

Number of Covariates

Between $q = 2$ and $q = 20$ covariates $\mathbf{x}$ were randomly included in each replication (random design factor D). Across replications, on average of eleven covariates were included in the model. The DIF detection included between 12 and 120 model parameters for intercepts ($\mathbf{K}$) and factor loadings ($\mathbf{\Omega}_1$), respectively.

Proportion of Continuous Covariates

Covariates were either standard normally distributed continuous variables with zero means and unit variances; or, they were binary dummy-coded variables representing group memberships (note that group memberships were allowed to be correlated and differences in variances were taken into account for model-generation and effect sizes). The proportion of continuous covariates in a replication was randomly sampled in range between 0% and 100% (random design factor E). Multivariate covariates were generated from a joint distribution (Demirtas & Doganay, 2012) using the `R` package `BinNor`.

Prevalance of DIF

Between 0% and 25% of the possible coefficients in the matrices $\mathbf{K}$ and/or $\mathbf{\Omega}_1$ contained DIF (random design factor F). If factor loadings and intercepts produced DIF, they were specified with the same pattern. The actual pattern was randomly chosen from all possible positions in the matrices. These conditions included the important situation where items showed barely any DIF or even full MI held as well as where several items showed DIF caused by several covariates. This second situation is challenging for applied researchers because DIF is caused by correlated variables in a complex pattern.

Multicollinearity

The correlations among covariates and of the covariates to the factor were randomly assigned to values between $\rho = 0$ and $\rho = 0.7$ using a uniform distribution $U(0, 0.7)$ (random design factor G).

This design resulted in four random design factors (D, E, F, G) and three fixed design factors with $3(A) \times 2(B) \times 2(C) = 12$ conditions for the data generation (between factors). Under each of the fixed design factor conditions $R = 1,000$ data sets were generated in `R` (i.e., 12,000 replications). All conditions are summarized in Table 2.

Data Analysis

Each of the data sets was repeatedly analyzed with each of the eight different priors (within factor). The specification of each prior is presented in Table 1. The first type of prior was a normal prior—$N(0, 1)$—that served as a baseline; its performance illustrates how the model performs under standard priors (i.e., without regularization). Note that this prior without the use of anchor items here is only very weakly identified. Second, in line with previous implementations for Bayesian DIF detection, we specified a small variance prior—$N(0, 0.001)$—based on the recommendations by Shi et al. (2017). Third, traditional lasso type Laplace priors were used in the form of a Bayesian lasso (Blasso) and an adaptive

Table 2
Simulation Conditions With Three Fixed and Four Random Design Factors

Factor	Levels			
(A) Sample size (fixed)	200	400	800	
(B) DIF type (fixed)	**K**	**K** and $\mathbf{\Omega}_1$		
(C) DIF size (fixed)	0.1	0.2		
(D) Number of covariates (random)	2	to	20	
(E) Proportion of continuous covariates (random)	0%	to	100%	
(F) DIF prevalence (random)	0%	to	25%	
(G) Correlations covariates (random)	0.0	to	0.7	
Priors	$N(0, 1)$ Horse	$N(0, 0.001)$ DL	Blasso *N*-spike	aBlasso δ-spike

Note. Each data set was analyzed with eight different prior sets. $N(0, 1)$ = normal prior; $N(0, 0.001)$ = small variance prior; Blasso = Bayesian Lasso; aBlasso = adaptive Baysian Lasso; Horse = horseshoe prior; DL = Dirchlet Laplace; *N*-spike = continuous spike-and-slab prior; δ-spike = categorical spike-and-slab prior; DIF = differential item functioning.

Bayesian lasso (aBlasso). Fourth, one-component type global–local priors were used: the horseshoe[3] and the DL prior. Fifth, two-component type spike-and-slab priors were specified in a continuous version with a mix of two normal distributions ("*N*-spike," Ishwaran & Rao, 2005) and a version with a mix of a point mass (δ_0) and a Laplace distribution ("δ-spike" Brandt et al., 2018). More information on the technical aspects of the prior specifications in the simulation study can be found in "Appendix A."

For data analysis, all models used a saturated model structure with regard to $\mathbf{\Omega}_1$, $\mathbf{K}$, $\mathbf{\Gamma}_1$, and $\boldsymbol{\beta}_{11}$. No anchor items were used and thus models were only approximately identified.

All models were implemented in `Jags` (Plummer, 2003). Three chains with 10,000 iterations each were run. The first 5,000 iterations were discarded as burn-in. Convergence was checked for each analysis by investigating the Rhat statistic. Only analyses with Rhat statistics below 1.1 for all parameters were used for subsequent analyses.

DIF in an item due to a specific covariate was assumed if the 95% credible interval (CI) for the respective element in $\mathbf{K}$ or $\mathbf{\Omega}_1$ did not include zero. We chose this criterion because it could be implemented for each prior and can be considered an objective criterion (for a discussion of alternative criteria for Bayesian spike-and-slab priors, see Chen et al., 2021). For the results, type I error rates as well as power rates were calculated based on the proportion of false positive and true positive DIF in intercepts and factor loadings, respectively.

Expectations

Based on the theoretical properties of the different priors, we have the following expectations: For normal priors, we expect a low convergence rate and a suboptimal performance due to the specification without anchor items. We expect that the small variance prior will have a lower power especially for smaller sample sizes because it uses a very strict prior without the flexibility of modern shrinkage priors. We further expect that all shrinkage priors will have acceptable type I error rates that are likely to be very conservative (clearly below the nominal level of 5%). With regard to power, it is difficult to make predictions about differential performance of the modern shrinkage priors because they have not been investigated for such a complex latent variable model yet. However, any differences will decrease with increasing sample size; they might be negligible with large sample sizes (e.g., $N = 800$). Finally, we expect that at least a sample size of $N = 400$ is necessary to achieve an acceptable level of 80% power to detect DIF.

Results

Here, we will first provide information about convergence rates, type I error rates and power focusing on the fixed design factors (sample size, size, and type of DIF). Then, we will explore the relationships with the random design factors (number and type of covariates, prevalence of DIF, and multicollinearity).

Convergence

Figure 1 shows the convergence rates under different conditions of the fixed design factors type and size of DIF, and sample size. Convergence rates were as low as 17.4% for normal priors, which was expected due to the missing anchor items. Convergence improved when DIF was large and part of both factor loadings and intercepts with a maximum of 70.4%. Even though this convergence rate was low, we kept the normal prior in the following discussion as a baseline (or, worst-case scenario). *N*-spike, aBlasso, and DL priors had lower convergence rates than the remaining shrinkage priors with values between 30.9% and 90.2% (with average convergence rates of 60.2%, 71.6%, and 73.2% for *N*-spike, aBlasso, and DL priors, respectively. For the remaining priors, convergence rates were above 99.8% (small variance prior), 89.4% (Blasso), 89.5% (horseshoe), and 91.9% (δ-spike).

Convergence rates (maximum Rhats per replication) were fairly unrelated to the random factors (e.g., number of covariates). Detailed results can be found in Figure B1 in "Appendix B."

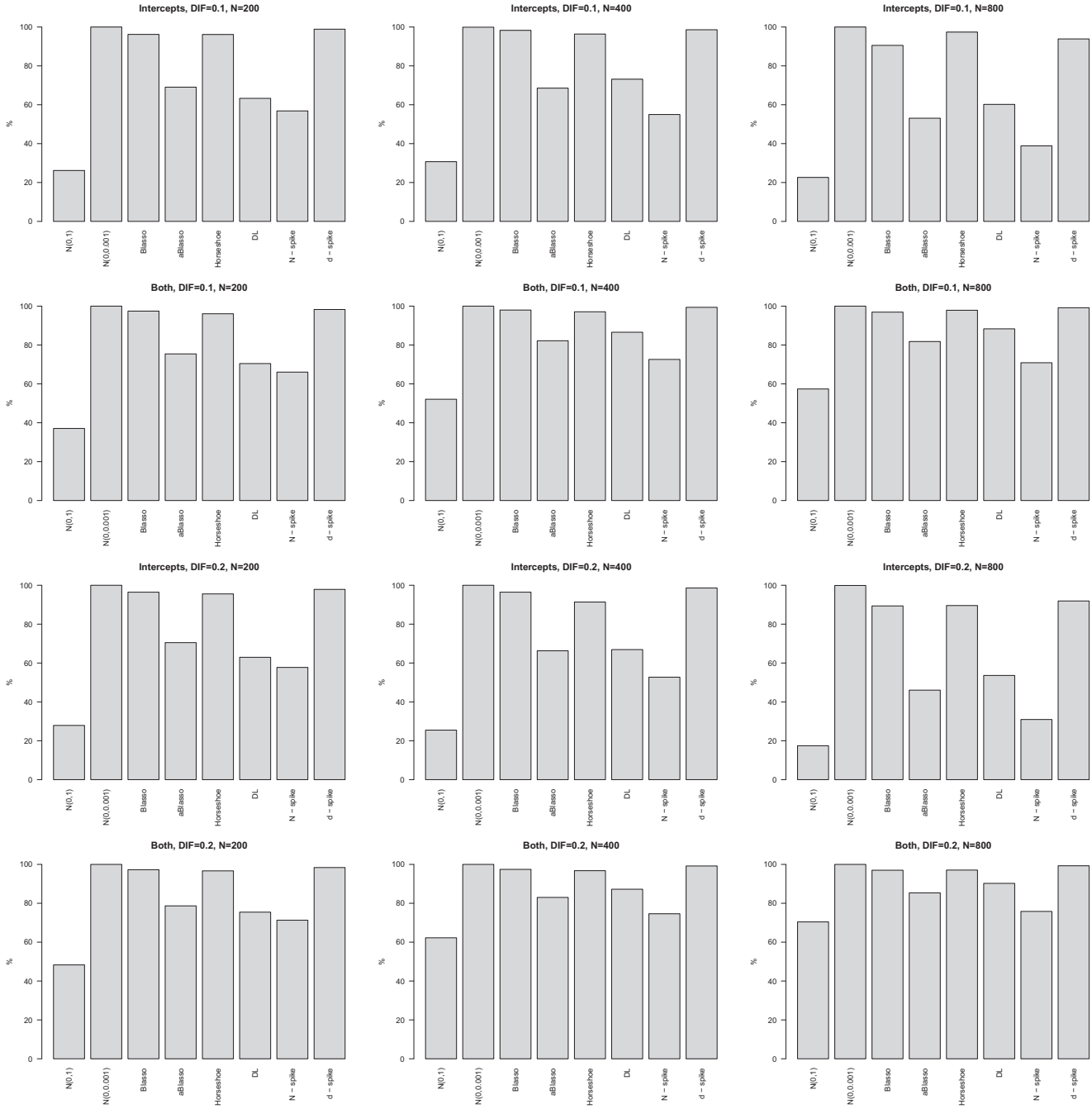
DIF in Intercepts

Figure 2 shows the false positive (type I error in black) and true positive (power in gray) rates under different conditions for the detection of DIF in the intercepts as superimposed bars. As reference lines, dashed lines indicate nominal levels of 5% type I error rate and 80% power; the dotted line shows the highest power achieved with any prior except the normal prior.

[3] We also included the horseshoe+ prior in the analyses, but the results were virtually identical to the horseshoe. We decided to skip the results.

Figure 1

Convergence Rates Under the Different Fixed Factor Design Conditions

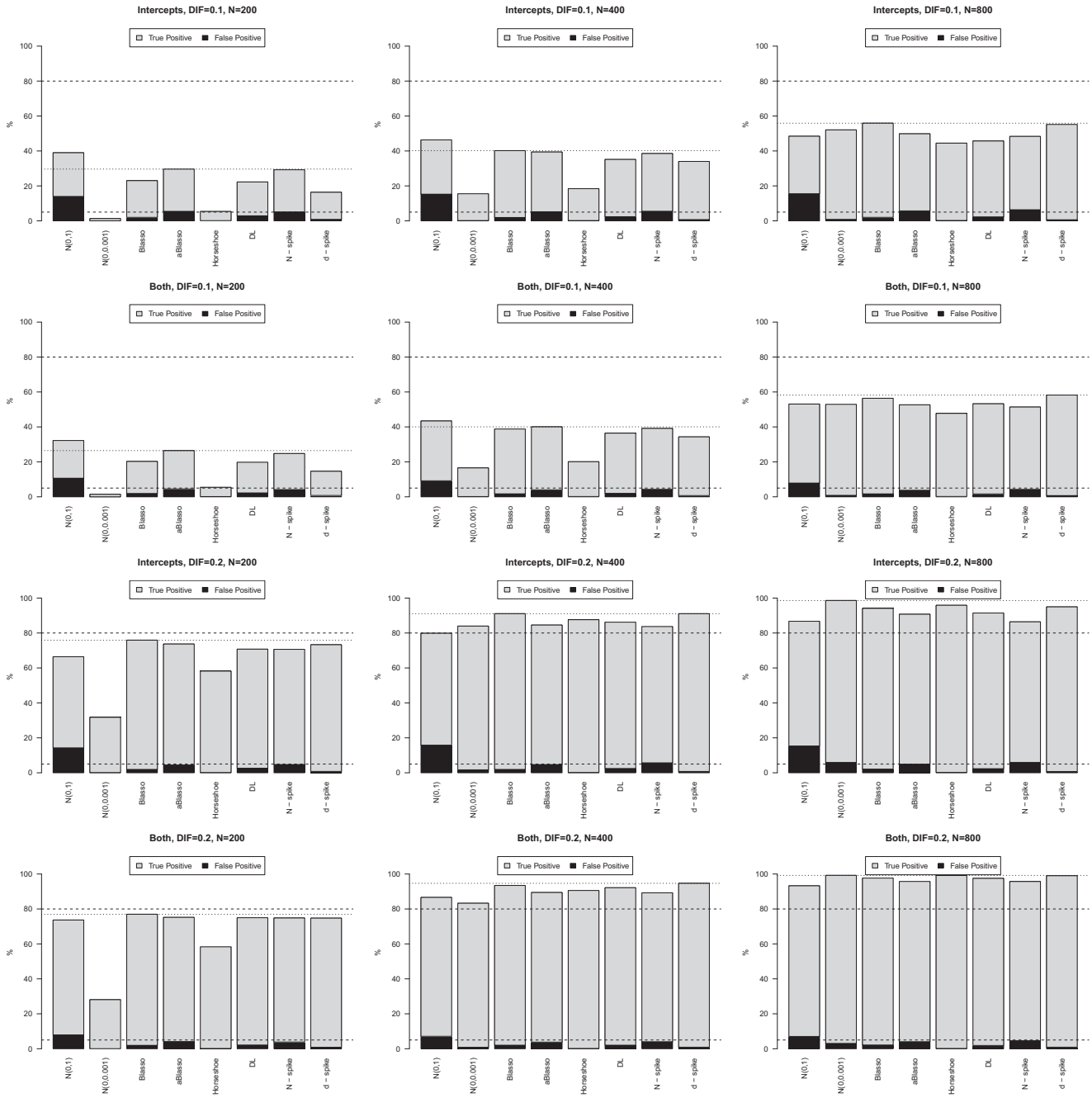


With regard to the false positive rate, only the normal prior showed inflated levels. These levels were higher when DIF was only present in the intercepts with rates above 15.8% compared to rates above 10.6% when DIF was present in both. All shrinkage priors showed acceptable levels. They were close to the nominal 5% level with maximum type I error rate of 6.3% (*N*-spike), or 5.7% (aBlasso). For the horseshoe, DL, Blasso, and δ-spike priors, they were close to zero with a maximum of 2.9% (for the DL prior). The small variance priors had close to zero type I error rates except under the condition of large sample size and large DIF in intercepts with a type I error rate of 6.0%.

Power (true positive rate) was affected strongly by the size of DIF and sample size; it was virtually identical for DIF in intercepts only versus DIF in intercepts and factor loadings. For sample sizes of $N = 400$ or $N = 800$ in combination with large DIF, results were fairly similar across all priors with values above 79.8% power (normal prior); the highest minimal power was provided by δ-spike and

Figure 2

False Positive (Type I Error) and True Positive (Power) Rates the Different Fixed Factor Design Conditions for the Detection of DIF in the Intercepts



Note. Dashed lines indicate nominal levels of 5% type I error rate and 80% power; the dotted line shows the highest percentage achieved with any prior except the normal prior. DIF = differential item functioning.

Blasso each with 91.1%. For small sample sizes of $N = 200$ in combination with large DIF, the aBlasso and Blasso, δ-spike and *N*-spike, as well as DL priors had similar minimal power rates with values between 70.6% and 75.9%. The horseshoe and the normal prior had lower power rates with minimal value of 58.2% and 66.4%, respectively. The small variance prior had a very low power with rates as low as 28.1%.

For $N = 200$ in combination with small DIF, the power for those methods that did not produce unacceptably inflated type I error rates (i.e., the normal prior) ranged between lower levels with 1.4% and 5.4% for the small variance prior and the horseshoe, respectively, and higher levels with 29.2% and 29.6% for the *N*-spike and aBlasso. For $N = 400$ and small DIF, all shrinkage priors except of small variance and horseshoe priors showed a similar

performance with an average power between 34.2% (δ-spike), and 36.7% (aBlasso). Finally for $N = 800$ and small DIF, these priors had a average power between 49.5% (DL) and 56.7% (δ-spike). Under both sample sizes, the horseshoe and small variance priors again showed lower levels of power.

DIF in Factor Loadings

False positive rates for the DIF in factor loadings were close to zero across all conditions for all priors except the normal prior with a maximum rate of 3.3%. Only for the normal prior, the type I rate was slightly inflated within a range of 4.9% and 8.9%.

The power (true positive rate) to detect DIF in factor loadings was generally lower than intercepts except for large sample sizes and large DIF. For small DIF, sample sizes needed to be large (i.e., $N = 800$) in order to provide at least a power of 47.3% (aBlasso) to 53.7% (small variance prior), although priors performed rather similarly under this sample size. For medium DIF, the power already reached a nominal 80% level at $N = 400$. Again priors provided fairly similar results for sample sizes like this or larger. For $N = 200$ and a large DIF, the small variance prior and the horseshoe showed comparatively lower power rates with 38.5% and 52.6%, respectively, whereas the remaining priors showed similar power rates between 58.7% (*N*-spike) and 65.4% (DL prior).TEST

Number of Covariates

Figure 4 shows the relationships between true and false positive rates for both intercepts and factor loadings with regard to the number of covariates in the model (summarized across the remaining design conditions). The inflated type I error rate for the intercept using normal priors was consistent across the number of covariates (top left panel). For some of the remaining priors, a positive relationship between type I error rate and number of covariates could be observed, particularly for the aBlasso and *N*-spike priors, but it remained below the nominal 5% level on average. For the false positive rates of the factor loadings, a similar pattern could be observed on a slightly lower level.

The power to detect DIF in the intercepts remained constant across the number of covariates for the small variance prior and the horseshoe. For the other shrinkage priors, a similar positive relationship could be observed. The power to detect DIF in factor loadings was independent of the number of covariates for the small variance prior, the horseshoe, the Blasso, and the δ-spike priors whereas it increased for the *N*-spike, aBlasso, normal, and DL priors.

Proportion of Continuous Covariates

Figure 5 shows the relationships between true and false positive rates with the proportion of continuous covariates in the model. The type I error rates were fairly constant across the proportion of continuous covariates for most priors. For the normal prior, a slightly higher rate was observed when the number of continuous and categorical covariates were equal, even though the overall impact was close to zero.

The power rates for DIF in intercepts were fairly independent of the percentage of continuous covariates across all priors. For the factor loadings, a slight advantage for increasing proportions of continuous covariates could be observed (particularly above 50%).

Prevalance of DIF

Figure 6 shows the relationships between true and false positive rates compared with the prevalence of DIF in the model (i.e., the proportion of intercepts and/or factor loadings with DIF). The false positive rate was fairly independent across the percentage of DIF. Only for intercepts, the observed the type I error rate inflation of the normal prior decreased with increasing DIF (i.e., with 5% or less DIF, type I rates were above 10%).

The power to detect DIF decreased with increasing percentage of DIF for all priors. Across conditions, the aBlasso and *N*-spike priors showed the best performance. Lowest power levels were observed for small variance and horseshoe priors.

Multicollinearity

Figure 7 shows the relationships between true and false positive rates with the degree of correlation among covariates and between covariates and the latent factor. For the DIF in intercepts, inflated type I error rates of the normal prior were more prominent when multicollinearity increased. The type I error rates for the DIF in factor loadings were fairly independent of the correlation.

The relative performance for the true positive (power) rates for the DIF in intercepts was fairly stable across different levels of multicollinearity for all priors except for the small variance and horseshoe priors. For these two priors, power went down with increasing multicollinearity starting at about $\rho = 0.3$.

The power to detect DIF in factor loadings was impacted negatively by the multicollinarity for all priors. For correlations between $\rho = 0$ and $\rho = 0.4$, the power was fairly constant on the respective power level of each prior. For correlations above $\rho = 0.4$, the power was reduced by up to about 20% at $\rho = 0.7$.

Relevance of Design Factors

Finally, Table 3 summarized the relative relevance of the design factors. For simplicity, we assumed no interactions between the factors. We calculated the incremental variance explained by each design factor (ΔR^2) for each prior and for each relevant dependent variable compared to a model that included all factors.

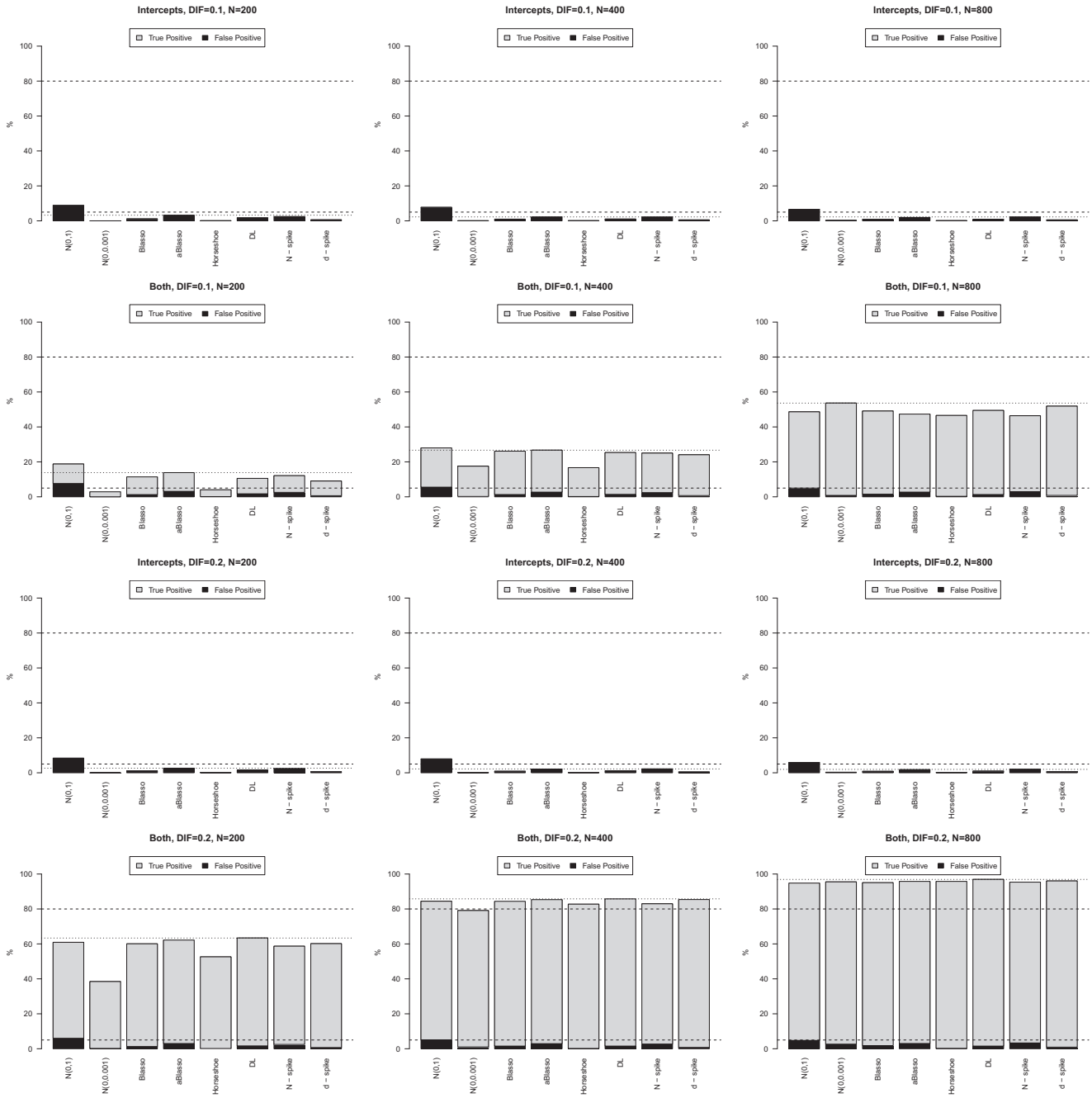
For the true positive rates (power), the most relevant factor was the size of DIF with ΔR^2s between 30.8% (DIF in intercepts for normal priors) and 43.7% (DIF in intercepts for the horseshoe). The second (obvious) relevant factor was the sample size, though it was more relevant for the factor loadings than for the intercepts.[4] For the factor loadings, the shrinkage priors showed a similar relevance with ΔR^2s between 11.4% and 17.5% compared to the normal prior with 9.2% and the small variance prior with 27.9%. For the DIF in intercepts, the pattern was more heterogeneous with a range between 3.2% (normal prior) and 34.9% (small variance prior).

For the false positive rates, we observed overall that the impact of the conditions were rather small, which is to be interpreted as

[4] Note that this finding needs to be interpreted relative to the chosen factor levels, that is, $N = 200, 400, 800$. A choice of other sample sizes would of course change the importance of this factor. The relevance of this finding is linked to the representativeness of the chosen levels.

Figure 3

False Positive (Type I Error) and True Positive (Power) Rates Under the Different Fixed Factor Design Conditions for the Detection of DIF in the Factor Loadings



Note. Dashed lines indicate nominal levels of 5% type I error rate and 80% power; the dotted line shows the highest percentage achieved with any prior except the normal prior. DIF = differential item functioning.

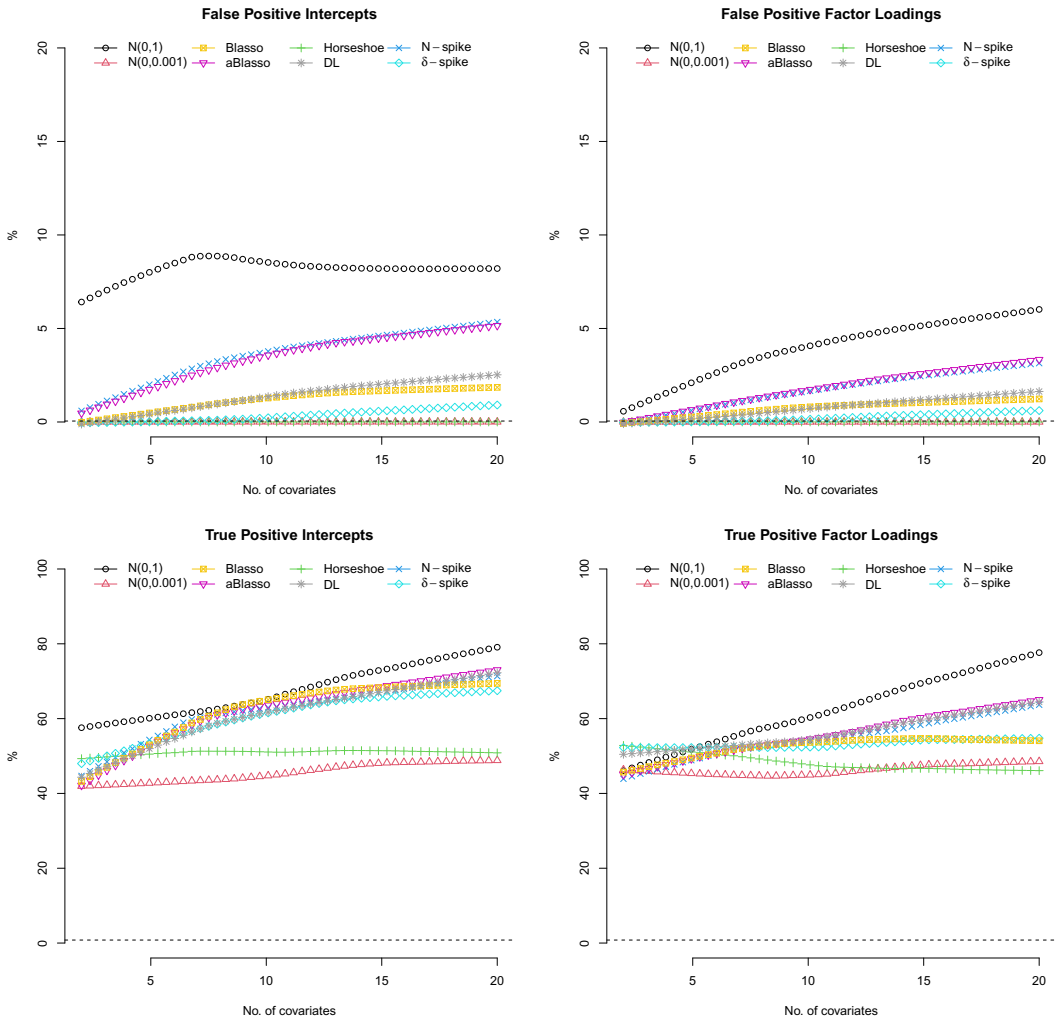
positive (because the type I error rates should not be affected by them). The only prior that showed an inflated type I error rate was the normal prior; the main impacts for DIF in intercepts were the type of DIF (in intercepts only vs. both) and the multicollinearity. For the factor loadings, the relevant factor for the normal prior was the number of covariates with 5.9% additional explained variance.

Summary of the Study

Summarizing the results, we would like to make the following observations: The normal prior showed suboptimal performance with low convergence rates and inflated type I error rates. The small variance prior provided very low power for smaller sample sizes compared to the other priors; only for larger sample sizes, this

Figure 4

Relationship Between False Positive (Type I Error) and True Positive (Power) Rates for the Detection of DIF in the Intercepts and Factor Loadings With Regard to the Number of Covariates (Loess Approximation)



Note. DIF = differential item functioning. See the online article for the color version of this figure

disadvantage was overcome. When sample size and DIF were large ($N = 800$) all shrinkage priors showed a similar performance. While all shrinkage priors kept the nominal type I error rate, the horseshoe prior often showed a slightly lower power to detect DIF.

Empirical Example

In this section, we illustrate the practical value and implementation of Bayesian MNLFA with regularizing priors within the context of a large-scale assessment study, the PISA 2018 study (OECD, 2019), which examined students interests and motivations across many different countries. The scope of DIF to consider for this application would make it difficult if not impossible to determine MI using more traditional methods. The PISA study is a large-scale assessment of students' achievements as well as interests and motivation across many different countries. Data are publicly available under https://www.oecd.org/pisa/data/2018database/.

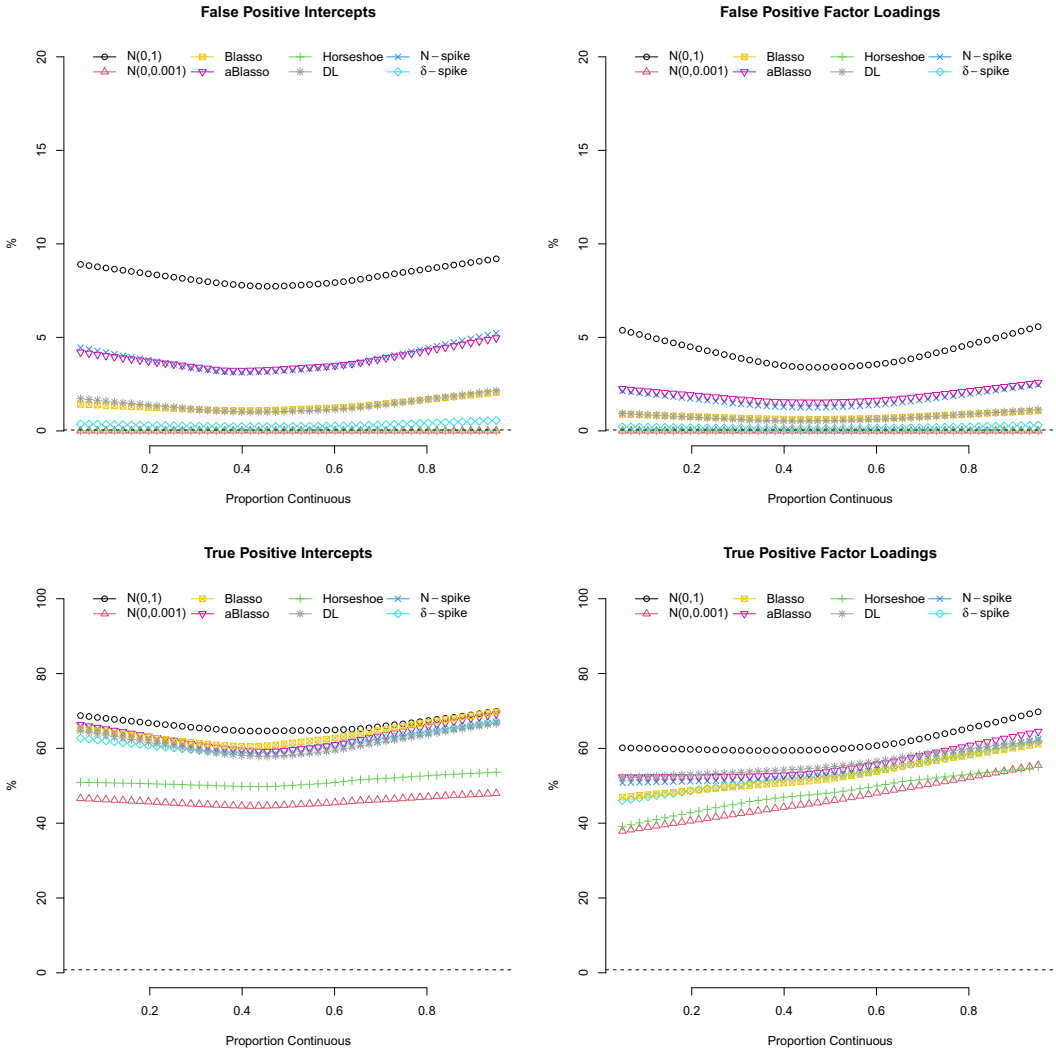
Data Preparation

Data were used only for the scale "Usefulness for understanding and memorizing text" (ST164) that was comprised of six items such as "I summarize the text in my own words" or "I read the text aloud to another person." Answers were given on a 6-point Likert scale ranging from "Not useful at all" to "Very useful." Here, we used data from a subset of nine countries (Belgium, Brazil, Switzerland, Germany, France, Greece, Italy, Turkey, and Ukraine). These countries were selected with the expectations that some of them should not show much DIF because they are similar to Germany (e.g., Belgium or France) whereas others that are less culturally similar might be expected to manifest more DIF (e.g., Brazil or Switzerland).

From each of the countries, data were available for 4,304 to 10,665 students. From these students, 100 were selected randomly for each country, so that the total sample size in this analysis was $N = 900$, which is comparable to the conditions in the simulation

Figure 5

Relationship Between False Positive (Type I Error) and True Positive (Power) Rates for the Detection of DIF in the Intercepts and Factor Loadings With Regard to Proportion of Continuous (vs. Binary) Covariates (Loess Approximation)



Note. DIF = differential item functioning. See the online article for the color version of this figure

study. Eight covariates were created by dummy coding the country with Germany as a reference group (i.e., the factor mean was set to zero and its variance to one). Six additional covariates were included: gender, age (mean of 16 with an *SD* of 0.28), availability of equipment for studying (e.g., desk, internet), availability of fine arts (e.g., books on art, music, or design), a proxy for SES (e.g., number of televisions, cars, and cell phones), and the number of books (a proxy for the literacy level of the family). All continuous covariates were standardized prior to analyses, and all dichotomous covariates were centered (weighted effects coding).

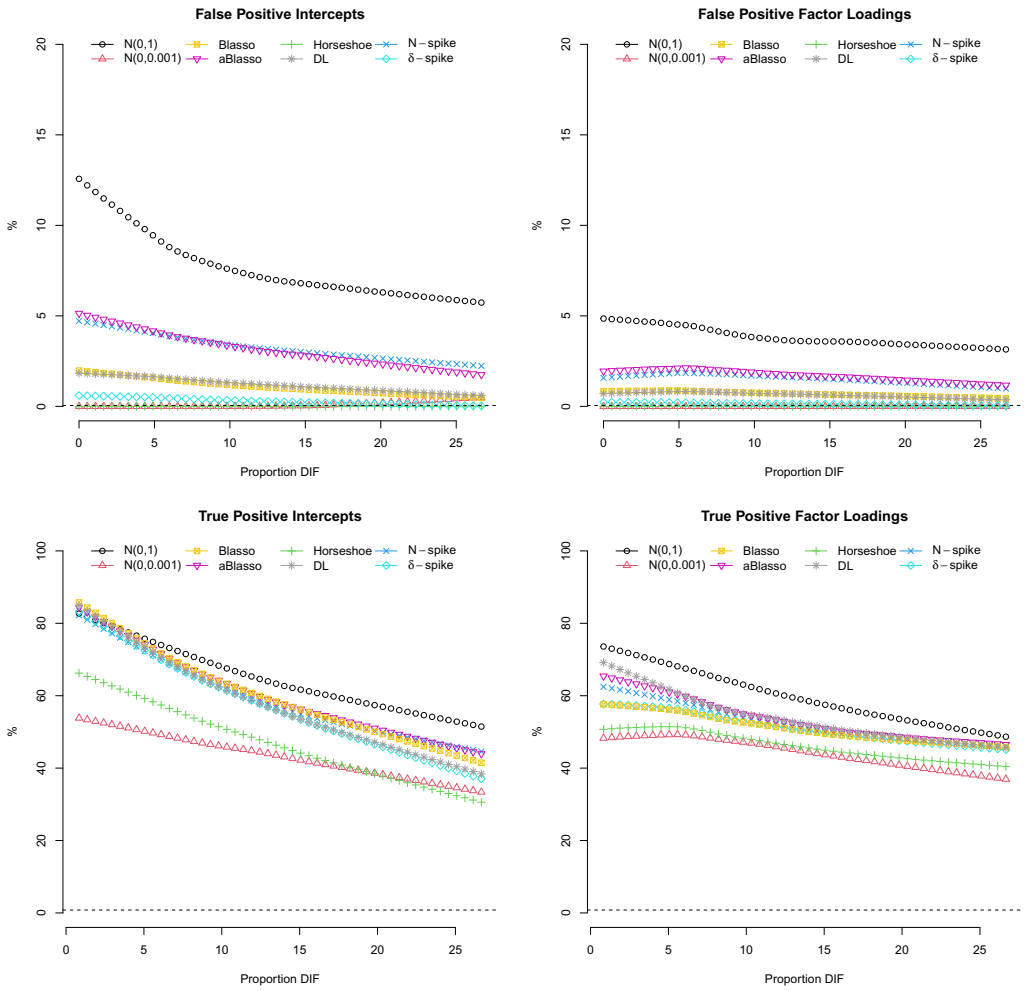
Data Analysis

For the analysis, all eight prior sets as described in the simulation section were applied. Based on the simulation study and the empirical characteristics of the data—such as low proportion continuous, $N = 900$, largely uncorrelated covariates (with countries being independent), etc.—we mostly trust the results from lasso type and spike-and-slab priors; the normal, small variance, or horseshoe priors are less trustworthy but they are included as a comparison (and illustration of their performance in empirical data). Analyses were conducted in `R` and `Jags` using identical hyperpriors as in the simulation.[5] Three chains each with 10,000 iterations were run; 5,000 iterations were discarded as burn-in. All models converged, which was checked graphically and using the Rhat statistic. The largest Rhat value was produced by the *N*-spike prior with 1.06. Trace

[5] The complete R script including the jags syntax can be found on the first author's website www.holger-brandts-methods.com.

Figure 6

Relationship Between False Positive (Type I Error) and True Positive (Power) Rates for the Detection of DIF in the Intercepts and Factor Loadings With Regard to the Prevalence of DIF (Loess Approximation)



Note. DIF = differential item functioning. See the online article for the color version of this figure

and density plots were inspected and indicated sufficient chain mixing.

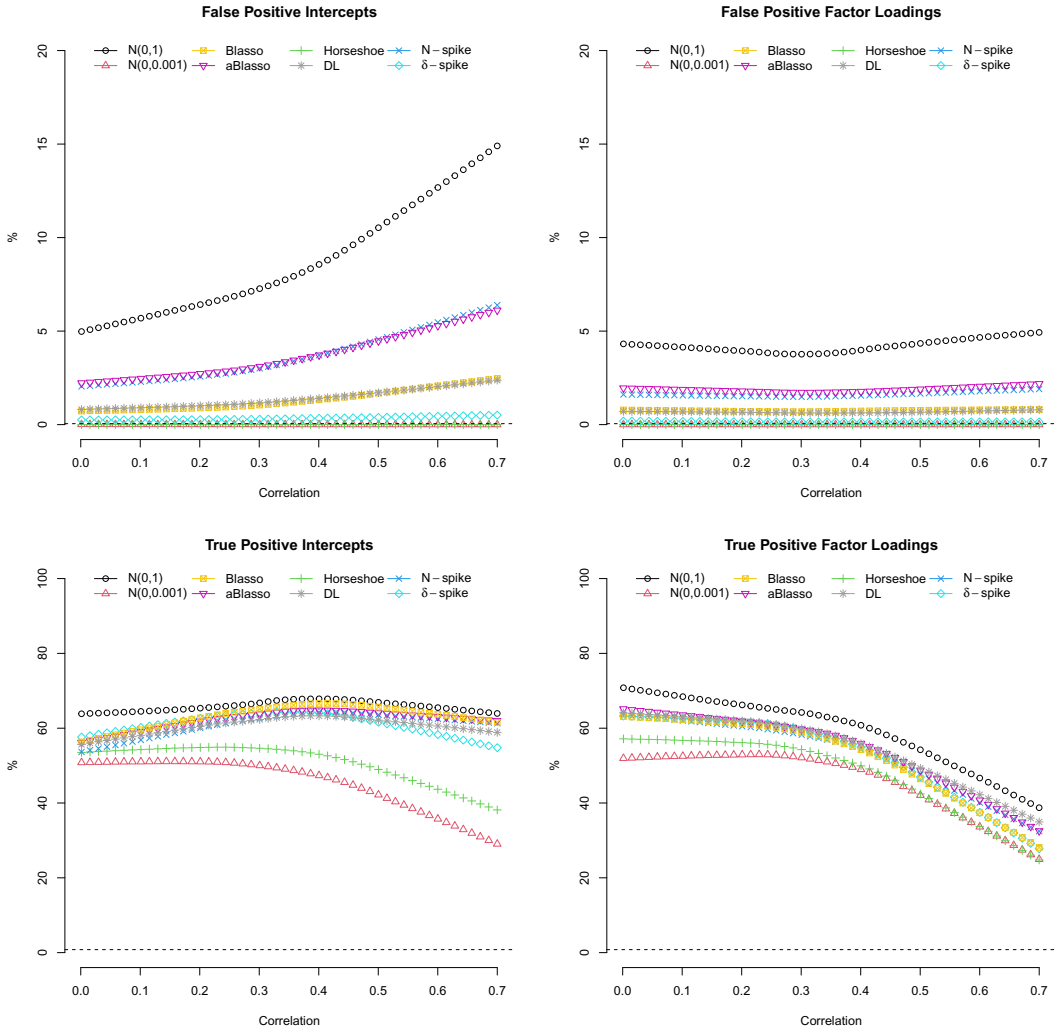
For the effective sample size (ESS), we calculated a minimum necessary multivariate ESS of 7,713 for the 215 model parameters (Vats et al., 2015). All priors except of the horseshoe prior provided a multivariate ESS above this threshold with values ranging between 7,970 (adaptive lasso) and 17,466 (small variance prior). The horseshoe's multivariate ESS reached acceptable levels with 5,000 additional iterations (multivariate ESS of 9,744 vs. 6,483). The results from both posteriors using the horseshoe were otherwise virtually identical.

In addition, we investigated the minimum univariate ESS, which was above the recommended guideline of an ESS above 400 (cf. Lüdtke et al., 2018; Zitzmann & Hecht, 2019) for the small variance and horseshoe priors. The remaining priors showed a univariate ESS between 150 (normal prior) and 342 (δ-spike). Increasing the chain length to 50,000 with 25,000 burn-in iterations improved the univariate ESS to a minimum of 714 (aBlasso) for all priors except for the normal prior (with an ESS of 295). At the same time, the posteriors of both runs with shorter and longer chain lengths were virtually identical for all priors except for the normal prior with an average squared differences of posterior means, 2.5% and 97.5% percentiles below 0.009, 0.064, and 0.083, respectively, between the two runs. For consistency with the simulation study, we will report the results with 10,000 iterations for all priors here. We included the results for the normal prior even though its performance was suboptimal because it illustrates the performance of a model without shrinkage.

We treated the 6-point Likert-type items as continuous, following recommendations from a number of simulation studies demonstrating that this results in minimal bias at the sample size here and may increase the stability of the estimates (Bandalos, 2014; Barendse et al., 2015; DiStefano, 2002; Lei & Shiverdecker, 2020; Liu & Thompson, 2020; Sass et al., 2014). We tested the normality of the residuals as well as the linearity assumption using factor and residual scores obtained from the analyses (see "Appendix C"). The *QQ* plots for the residuals indicated fairly normal distributions for all

Figure 7

Relationship Between False Positive (Type I Error) and True Positive (Power) Rates for the Detection of DIF in the Intercepts and Factor Loadings With Regard to Correlation Between Covariates (Loess Approximation)



Note. DIF = differential item functioning. See the online article for the color version of this figure

items with some rather minor deviations from normality for Y_1, Y_2, Y_4, and Y_6. The partial plots using loess approximations for the relationship between factor scores and items indicated fairly linear functions for Y_3 and Y_4, whereas the remaining items showed some slight deviation from linearity (particularly Y_1, Y_2, and Y_6). The violations of the assumptions were deemed not too severe to affect the interpretations with regard to the DIF detection in the followings subsections.[6]

Results

Estimates for the factor loadings λ_0 (here, for German, female students with average scores in the continuous covariates) were very similar across priors and they were all significant (i.e., the 95% CI did not include zero). On average, they were small for items 1 and 2 ($\bar{\lambda}_{01} = 0.36$ and$\bar{\lambda}_{02} = 0.30$) and higher for the remaining four items (between $\bar{\lambda}_{06} = 0.82$ and $\bar{\lambda}_{05} = 1.10$). Item intercepts were also very similar across priors and they were all significant (with estimates lying between 2.94 and 4.44).

Figure 8 illustrates the parameter estimates for the DIF in intercepts and factor loadings for the first item "I concentrate on the parts of the text that are easy to understand" (the plots for the remaining five items are depicted in "Appendix C").

For the intercepts, DIF was observed for Greece (normal and aBlasso), Switzerland (normal, aBlasso, DL, and *N*-spike), and Brazil (normal, aBlasso, and DL). This implied that these countries might show larger intercepts compared to Germany (i.e., students from these countries agreed more often with the item). However,

[6] While some work on DIF-detection for binary data has been conducted (Chen et al., 2021), an extension to ordinal data is not straightforward. Different parameterization options of DIF may be considered for graded response models, but an optimal selection needs further research that goes beyond what we can investigate in this article.

Table 3

Percent Incremental Variance Explanation (ΔR^2) for False and True Positive Rates (Type I Error and Power) for Intercepts Factor Loadings Based on the Seven Design Factors, Separate for Each of the Prior Sets

Prior	*N*	*type*	*size*	*No. Cov*	*%cont.*	*%DIF*	ρ
False positive rates DIF intercepts							
N(0,1)	0.0	**5.9**	0.7	0.5	0.0	2.1	**5.0**
N(0,0.001)	**9.2**	0.3	**9.4**	0.0	0.0	**11.8**	0.8
Blasso	0.0	0.1	0.0	0.4	1.3	1.0	**7.7**
aBlasso	0.1	1.3	0.5	2.8	0.4	1.6	**6.8**
Horse	0.1	0.0	0.1	0.3	0.0	0.9	0.1
DL	0.7	1.0	0.0	**8.4**	0.4	0.5	**9.0**
N-spike	0.2	1.6	0.0	2.2	0.3	0.2	**8.6**
δ-spike	0.1	0.0	0.0	4.5	1.0	0.7	2.9
False positive rates DIF factor loadings							
N(0,1)	2.6	1.5	0.3	**6.0**	0.0	0.4	0.6
N(0,0.001)	4.8	4.4	3.3	0.0	0.1	4.6	0.6
Blasso	0.1	1.8	0.0	1.0	0.0	0.0	0.5
aBlasso	0.8	0.3	0.1	**8.2**	0.0	0.4	0.3
Horse	0.2	0.2	0.0	1.5	0.1	0.0	0.0
DL	1.3	0.2	0.0	**6.6**	0.0	0.1	1.3
N-spike	0.1	0.1	0.0	**6.7**	0.0	0.0	0.7
δ-spike	0.1	0.7	0.0	0.9	0.1	0.0	0.1
True positive rates DIF intercepts							
N(0,1)	3.2	0.1	**30.8**	2.1	0.0	0.4	0.9
N(0,0.001)	**34.9**	0.0	**29.9**	0.2	0.0	0.0	1.1
Blasso	**9.2**	0.0	**40.8**	2.7	0.2	0.0	0.6
aBlasso	**5.3**	0.0	**38.0**	3.8	0.0	0.1	1.2
Horse	**16.1**	0.0	**43.7**	0.0	0.0	0.1	0.8
DL	**6.8**	0.1	**42.1**	2.7	0.0	0.0	0.5
N-spike	**5.3**	0.0	**37.2**	2.6	0.0	0.1	1.2
δ-spike	**12.1**	0.0	**42.0**	1.4	0.1	0.2	0.0
True positive rates DIF factor loadings							
N(0,1)	**9.2**	NA	**39.1**	1.8	0.2	1.2	0.8
N(0,0.001)	**27.9**	NA	**29.4**	0.2	1.0	0.5	1.3
Blasso	**13.3**	NA	**38.3**	0.6	0.6	0.9	2.5
aBlasso	**11.4**	NA	**42.1**	1.3	0.4	1.3	1.4
Horse	**17.5**	NA	**39.3**	0.0	0.8	0.4	2.3
DL	**13.3**	NA	**43.1**	0.5	0.3	0.8	1.5
N-spike	**12.5**	NA	**40.1**	1.0	0.4	1.2	1.3
δ-spike	**15.7**	NA	**38.6**	0.1	0.8	0.6	2.8

Note. Dominant values above 5% are bold typed. *N* = sample size. *type* = type of DIF. *size* = size of DIF. *No. Cov* = number of covariates. *%cont.* =proportion of continuous covariates. *%DIF* = prevalence of DIF. ρ = multicollinearity. *N*(0, 1) = normal prior. *N*(0, 0.001) = small variance prior. Blasso= Bayesian Lasso. aBlasso = adaptive Baysian Lasso. Horse = horseshoe prior. DL = Dirchlet Laplace. *N*-spike = continuous spike-and-slab prior. δ-spike = categorical spike-and-slab prior. NA = not applicable. DIF = differential item functioning; DL = Dirchlet Laplace.

small variance, Blasso, horseshoe, and δ-spike priors did not indicate any DIF in the intercepts of Item 1.

For the factor loadings, DIF was identified consistently for SES and Fine Arts by all priors except of the horseshoe and small variance priors. For gender, DIF was detected by the normal prior, aBlasso, and DL. This implied that students with increasing SES or scores in Fine Arts as well as female students may have lower factor loadings.

From a practical perspective, this implies the following aspects: First, given the simulation results, results from the normal prior might not be reliable because of the type I error inflation. As a consequence, it seems reasonable to assume that the item has very similar intercepts across the covariates tested here. Primarily SES and Fine Arts are covariates that need to be taken into account for any subsequent analysis of the underlying constructs.

Table 4 summarized the results for the remaining five items. For the intercepts, results were very fairly similar with more significant results for the normal (14.3%) and aBlasso priors (13.1%), some significant results for the δ-spike, *N*-spike, Blasso, and DL priors (between 3.6% and 9.5%), and almost no DIF for small variance, and horseshoe priors (between 0% and 2.4%).

The overall amount of DIF in factor loadings across the 14 covariates and six items was similar for all priors and lay between 7.1% and 10.7%, except of the small variance and horseshoe priors (both 0%).

The majority of priors agreed on DIF for factor loadings in item y_2 ("I quickly read through the text twice.") caused by gender, Fine Arts, and Turkey. A similar agreement could be found for DIF in the factor loading of item y_6 ("I read the text aloud to another person.") due to SES, and for DIF in the intercept of item y_4 ("I underline important parts of the text.") due to gender. The remaining DIF's were identified predominantly by normal, aBlasso, and DL priors but not the others.

For the use of the scale "Usefulness for understanding and memorizing text," this implies that all items may have DIF in the intercept or the factor loading. Again, referring to the simulation study, particularly Items 2 and 4 show a consistent pattern that needs to be taken into account when investigating gender differences. Item 6 has a stronger relationship with SES.

How to Report DIF in Applied Settings

From a practical perspective, we recommend reporting similar tables when examining DIF for such complex scenarios. For example, Table 5 illustrates the results for the δ-spike prior. The table includes relevant information that allows researchers to judge how strongly MI is violated due to individual covariates. Two items did not show any DIF (y_3 and y_5). Both Fine Arts and SES affected the factor loading of y_1. While persons with an average score in these covariates had a factor loading of 0.35, persons with 1 *SD* above or below the average in Fine Arts (but with average SES) showed factor loadings of 0.20 and 0.50, respectively. This implied that (holding all remaining relevant covariates constant) the item "I concentrate on the parts of the text that are easy to understand" becomes less important for the factor score with increasing scores in Fine Arts. Similarly, item y_2 had a factor loading of 0.82 in Turkey compared to other countries with a factor loading of 0.29 (for persons with the same scores in the remaining covariates).

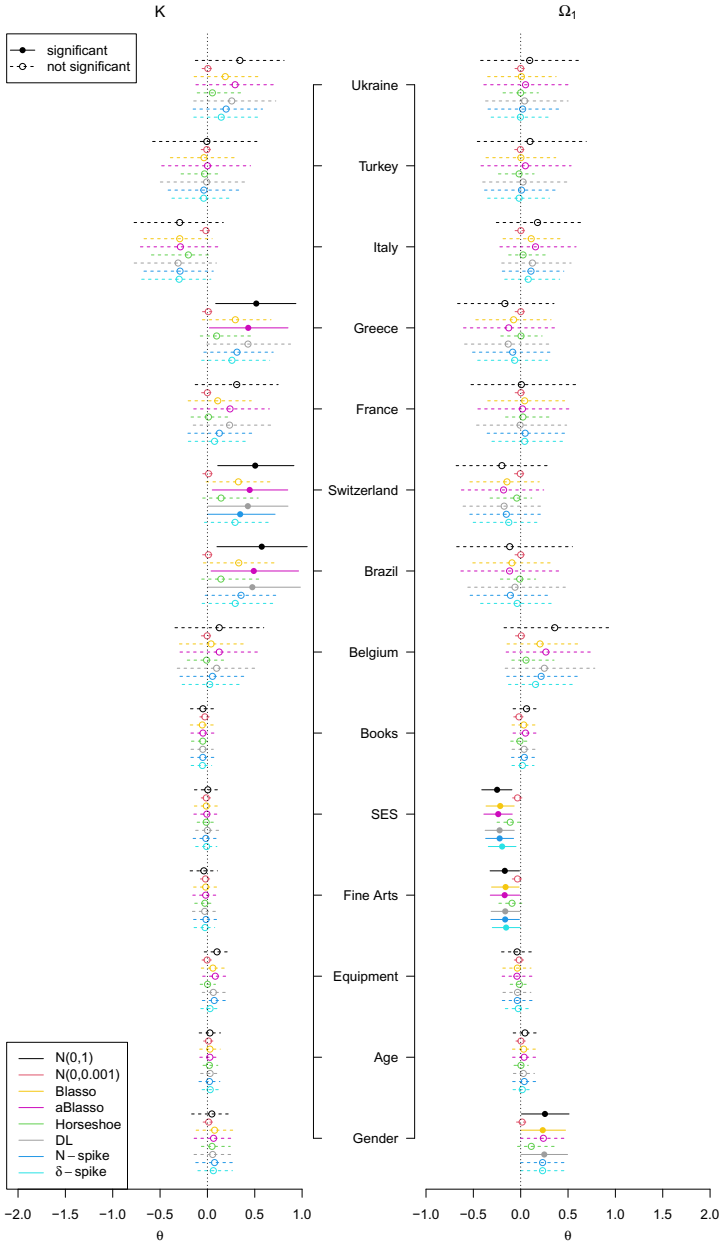
For the intercepts, Gender affected items y_2 and y_4, and SES affected item y_6. Hence these need to be taken into account when investigating differences in the underlying scale.

For a final comparison, we will test the hypotheses that male and female students have the same average score in the scale "Usefulness for understanding and memorizing text," and we test if SES affects the outcome in this scale. We will use as an example two models that are set up as posthoc analyses: The first model will not include any DIF to the covariates (baseline model). The second model will be based on the DIF detected with the δ-spike prior; standard normal priors will be used and DIF is only included for those coefficients that we identified with the δ-spike prior.

For the mean differences, the baseline model showed substantive difference for gender with a mean posterior estimate of − 0.28 and a

Figure 8

*Parameter Estimates and Credible Intervals for the DIF in Intercepts (**K**) and Factor Loadings (Ω_1) in the First Item ("I Concentrate on the Parts of the Text That are Easy to Understand") Across the 14 Covariates*



Note. Significant estimates are depicted with solid lines and filled dots. DIF = differential item functioning; SES = socio-economic status. See the online article for the color version of this figure.

95% credible interval of (− 0.45, − 0.11), but not for SES with an estimate of 0.06 (− 0.03, 0.15). If DIF is taken into account using model 2, both covariates show relevant effects with CI's of (− 0.53, − 0.16) for gender and (0.11, 0.30) for SES.

For the differences in the variances, the baseline model showed no substantive difference for gender (− 0.47, 0.15), but for SES (0.25, 0.56). For model 2, the factor variance did not depend on the covariates with CI's of (− 0.32, 0.32) and (− 0.13, 0.28). Taken together, one can conclude that female students rate the scale higher than male students with a similar spread in the responses. And, students with higher SES also reported higher scores in the scale compared to students with lower SES.

Table 4
Summarized Results for Items 2–5 With Regard to DIF in Intercepts and Factor Loadings

Item	Covariate	$N(0, 1)$	$N(0, 0.001)$	Blasso	aBlasso	Horse	DL	N-spike	δ-spike
				DIF in factor loadings					
y_2	Gender	1	0	1	1	0	1	1	0
	Fine Arts	1	0	1	1	0	1	1	1
	SES	1	0	0	0	0	0	0	0
	Turkey	1	0	1	1	0	1	1	1
y_3	Greece	1	0	0	0	0	0	0	0
y_5	Fine Arts	0	0	0	0	0	0	1	0
y_6	SES	1	0	1	1	0	0	1	1
				DIF in intercepts					
y_2	Gender	1	0	1	1	1	1	1	1
	France	1	0	0	0	0	0	0	0
	Italy	1	0	0	1	0	1	0	0
y_3	Greece	0	0	1	1	0	0	1	0
y_4	Gender	1	0	1	1	0	1	1	1
	Italy	1	0	0	1	0	1	0	0
y_5	Equipment	1	0	0	0	0	0	0	0
	France	1	0	0	1	0	1	0	0
	Greece	1	0	0	1	0	0	0	0
y_6	SES	1	0	1	1	1	1	1	1

Note. 1 indicates that DIF was present, 0 indicates no DIF. Items and covariates that did not show DIF for any prior were omitted in the table. Detailed results can be found in "Appendix C." $N(0, 1)$ = normal prior. $N(0, 0.001)$ = small variance prior. Blasso = Bayesian Lasso. aBlasso = adaptive Baysian Lasso. Horse= horseshoe prior. DL = Dirchlet Laplace. N-spike = continuous spike-and-slab prior. δ-spike = categorical spike-and-slab prior. DIF = differential item functioning.

Summary of the Empirical Example

Summarizing the results, at least two observations are necessary. First, from a methodological perspective, the performance of the priors is in line with what we would have expected from the simulation study. For example, the small variance priors showed close to zero estimates for all DIFs whereas the normal prior detected many (potentially inflated numbers of) relevant DIFs. For both the intercepts and factor loadings, all shrinkage priors showed a consistent pattern where the relevance of DIF (i.e., the CI did not include zero) was more overlapping for the factor loadings.

From an applied perspective, the pattern of DIF showed that both the magnitude and the prevalence of DIF should cause some concern when using scores from this scale in order to make comparisons, for example, across countries, gender comparisons, or even comparisons when persons stem from different levels of SES. It is mandatory that the identified DIF pattern is included in subsequent analyses as illustrated in the last subsection.

Table 5
Summarized Results for the δ-Spike Prior

Parameter	y_1	y_2	y_3	y_4	y_5	y_6
			Factor loadings			
λ_{10}	0.35	0.29	0.94	1.04	1.09	0.82
Fine Arts	−0.15	−0.19	—	—	—	—
SES	−0.20	—	—	—	—	−0.15
Turkey	—	0.53	—	—	—	—
			Intercepts			
ν_0	3.34	3.01	3.54	4.44	4.38	2.98
Gender	—	0.26	—	−0.38	—	—
SES	—	—	—	—	—	−0.17

Note. λ_{10} and ν_0 are the estimated baseline factor loadings and intercepts, respectively, the remaining coefficients indicate the size of DIF if the credible interval did not include zero (i.e., relevant DIF). Again, covariates that did not produce any DIF were omitted from the table. SES = socio-economic status; DIF = differential item functioning.

Discussion

In this article, we presented a Bayesian implementation of the MNLFA. We used different types of shrinkage priors in order to overcome limitations of traditional methods to detect DIF in situations with many (potentially correlated) covariates. We provided a detailed description of the different priors and how they can be implemented in MNLFA. In a simulation study, we investigated their performance. In an empirical example from educational psychology, we illustrated how this approach could be used to detect DIF. We now summarize our primary results and recommendations.

Guidelines and Recommendations

The simulation study used typical conditions researchers may encounter when analyzing their data. Overall, there were five important guidelines from this study.

First, we used a normal prior as a baseline because researchers might use them as they are standard for Bayesian models. In many situations, the normal prior produced inflated type I error rates. At the same time, it provided lower power and convergence rates than some of the shrinkage priors. We thus recommend that this prior be avoided when evaluating DIF.

Second, the small variance prior did not perform well when sample sizes were below $N = 800$. In these scenarios, the prior heavily impacted the sensitivity to detect DIF and produced power rates close to zero. The guidelines provided by Shi et al. (2017) should be viewed with caution in the MNLFA context. This result was

expected as the prior does not have the characteristics, for example, of spike-and-slab priors that allow single parameters to escape shrinkage via the slab. Instead, the small variance prior can be viewed as a spike-prior without the slab. This makes the prior too conservative under many conditions with smaller sample sizes.

Third, the horseshoe prior did not perform well and produced a lower power under most conditions compared to other shrinkage priors. One of the reasons might be found in the larger multicollinearity compared to previous simulation studies that showed a better performance of the prior but restricted scenarios to small or no multicollinearity (e.g., Bhadra et al., 2017).

Fourth, spike-and-slab and lasso priors showed similar performances under many conditions. Any of the priors will provide a good performance with regard to type I error rate and power under most conditions. From the two lasso priors, the non-adaptive version performed often slightly better than the adaptive version. For the spike-and-slab priors, the version with a delta spike at zero often had a slight advantage (e.g., with regard to convergence rates). Finally, as it is expected for Bayesian analyses, increasing sample sizes will lead to a similar performance of all priors. Hence, with increasing sample sizes, particularly above 800, performance differences due to a choice of a specific shrinkage prior will decrease.

Limitations and Future Perspectives

Simulation studies using complex models like the Bayesian MNLFA are necessarily limited in its amount of conditions that can be included. Computational costs of these models are high and running times are long. Here, we focused on several important conditions in which each of the priors will provide better or worse performances.

To be specific, we did not include larger number of covariates because in light of the findings, it would have been necessary to also include larger sample sizes. The number of items was fixed to six because this seems to be a typical length for a scale in psychological settings. It can be expected that larger sample sizes are necessary if scales include more items (like 20) but they are not as often used in applied settings.

We also followed recommendations by Bauer (2017) to investigate uni-dimensional scales only. If models with, for example, two factors are investigated, the model complexity increases dramatically. In this case, DIF for the factor loadings could occur both for the primary loadings but also for the potential cross-loadings. Additional work needs to be conducted to investigate these scenarios.

With regard to the priors, it should be noted that each of the priors comes at the cost of a sometimes difficult specification of hyperpriors. It was beyond the scope of this investigation to find optimal settings for each individual prior; instead we based our specification on recommendations by the original authors.

Finally, we did not tackle the aspect of model fit. Even though Bayesian model fit measures were recently proposed (e.g., Garnier-Villarreal & Jorgensen, 2020; Merkle et al., 2019), there are several limitations that need to be solved before a routine application to models with shrinkage priors can be used. The MNLFA is a nonlinear model for which standard fit indices do not directly apply (see discussion in Brandt et al., 2020). In addition, model fit in Bayesian SEM is tightly connected to prior choices because, for example, the degrees of freedom depend on how informative priors are (e.g., Garnier-Villarreal & Jorgensen, 2020). So far, limited information is available on the performance of Bayesian model fit measures and this is restricted to simple situations (informative vs. uninformative priors). No work exists for more complex situations like the shrinkage priors.

Notwithstanding the limitations noted above, we believe the current study demonstrates that the application of Bayesian shrinkage priors within MNLFA models represents both a theoretically compelling and empirically useful way to manage the complexity of evaluating DIF across many covariates simultaneously, especially in the absence of a priori known anchor items.

The material in this article was not presented at any conference or was disseminated prior to this presentation.

References

Armagan, A., Dunson, D. B., & Clyde, M. (2011). Generalized Beta mixtures of Gaussians. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira & K.Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 4, pp. 523–531). Curran Associates.

Armagan, A., Dunson, D. B., & Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, *23*(1), 119–143. PMID: 24478567.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling*, *21*(1), 102–116. https://doi.org/10.1080/10705511.2014.859510

Barendse, M. T., Oort, F., & Timmerman, M. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling*, *22*(1), 87–102. https://doi.org/10.1080/10705511.2014.934850

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform DIF: A simulation study. *Advances in Statistical Analysis*, *94*(2), 117–127. https://doi.org/10.1007/s10182-010-0126-1

Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2012). DIF detection through factor analysis. *Structural Equation Modeling*, *19*(4), 561–579. https://doi.org/10.1080/10705511.2012.713261

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling*, *27*(1), 43–55. https://doi.org/10.1080/10705511.2019.1642754

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101–125. https://doi.org/10.1037/a0015583

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673–690. https://doi.org/10.1037/met0000253

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals the horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, *12*(4), 1105–1131. https://doi.org/10.1214/16-BA1028

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*(512), 1479–1490. https://doi.org/10.1080/01621459.2014.960967

Bohrnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *American Statistical Association Journal*, *64*(328), 1439–1442. https://doi.org/10.1080/01621459.1969.10501069

Brandt, H., Cambria, J., & Kelava, A. (2018). An adaptive Bayesian Lasso approach with spike-and-slab priors to identify linear and interaction effects in structural equation models. *Structural Equation Modeling*, *25*(6), 946–960. https://doi.org/10.1080/10705511.2018.1474114

Brandt, H., Umbach, N., Kelava, A., & Bollen, K. A. (2020). Comparing estimators for latent interaction models under structural and distributional misspecifications. *Psychological Methods*, *25*(3), 321–345. https://doi.org/10.1037/met0000231

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). *Handling sparsity via the Horseshoe*. Proceedings of the 12th international conference on artificial intelligence and statistics (AISTATS) (Vol. 5, pp. 73–80).

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. https://doi.org/10.1093/biomet/asq017

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2021). Advantages of spike and slab priors for detecting differential item functioning relative to other Bayesian regularizing priors and frequentist lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(1), 122–139. https://doi.org/10.1080/10705511.2021.1948335

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, A. W., & Hussong, A. M. (2018). Recovering predictor-criterion relations using covariate-informed factor score estimation. *Structural Equation Modeling*, *25*(6), 860–875. https://doi.org/10.1080/10705511.2018.1473773

Demirtas, H., & Doganay, B. (2012). Simultaneous generation of binary and normal data with specified marginal and association structures. *Journal of Biopharmaceutical Statistics*, *22*(2), 223–236. https://doi.org/10.1080/10543406.2010.521874

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265–282. j.2044-8317.1992.tb00992.x

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*(3), 327–346. https://doi.org/10.1207/S15328007SEM0903_2

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Feng, X. N., Wang, G. C., Wang, Y. F., & Song, X. Y. (2015). Structure detection of semiparametric structural equation models with Bayesian adaptive group lasso. *Statistics in Medicine*, *34*(9), 1527–1547. https://doi.org/10.1002/sim.6410

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian SEM: Comparison to maximum likelihood. *Psychological Methods*, *25*(1), 46–70. https://doi.org/10.1037/met0000224

George, E. I., & Culloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

Griffin, J. E., & Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, *5*(1), 171–188. https://doi.org/10.1214/10-BA507

Guo, R., Zhu, H., Chow, S. M., & Ibrahim, J. G. (2012). Bayesian Lasso for semiparametric structural equation models. *Biometrics*, *68*(2), 567–577. https://doi.org/10.1111/j.1541-0420.2012.01751.x

Hans, C. (2009). Bayesian Lasso regression. *Biometrika*, *96*(4), 835–845. https://doi.org/10.1093/biomet/asp047

Hans, C. (2010). Model uncertainty and variable selection in Bayesian Lasso regression. *Statistics and Computing*, *20*(2), 221–229. https://doi.org/10.1007/s11222-009-9160-9

Huang, P. H. (2020). lslx: Semi-confirmatory structural equation modeling via penalized likelihood. *Journal of Statistical Software*, *93*(7), 1–37. https://doi.org/10.18637/jss.v093.i07

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773. https://doi.org/10.1214/009053604000001147

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Lei, P. W., & Shiverdecker, L. K. (2020). Performance of estimators for confirmatory factor analysis of ordinal variables with missing data. *Structural Equation Modeling*, *27*(4), 584–601. https://doi.org/10.1080/10705511.2019.1680292

Leng, C., Tran, M. N., & Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, *66*(2), 221–244. https://doi.org/10.1007/s10463-013-0429-6

Liu, Y., & Thompson, M. S. (2020). General factor mean difference estimation in bifactor models with ordinal data. *Structural Equation Modeling*, *28*(3), 423–439. https://doi.org/10.1080/10705511.2020.1833732

Lüdtke, O., Robitzsch, A., & Wagner, J. (2018). More stable estimation of the STARTS model: A Bayesian approach using Markov chain Monte Carlo techniques. *Psychological Methods*, *23*(3), 570–593. https://doi.org/10.1037/met0000155

Lykou, A., & Ntzoufras, I. (2013). On Bayesian Lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing*, *23*(3), 361–390. https://doi.org/10.1007/s11222-012-9316-x

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional vs. marginal likelihoods. *Psychometrika*, *84*(3), 802–829. https://doi.org/10.1007/s11336-019-09679-0

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032. https://doi.org/10.1080/01621459.1988.10478694

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/BF02296397

OECD. (2019). *Pisa 2018 results*(Vol. i). https://doi.org/10.1787/5f07c754-en.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*(2), 150–160.

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051. https://doi.org/10.1214/17-EJS1337SI

Plummer, M. (2003, March 20–22). *JAGS: A program for analysis of Bayesian graphical models using gibbs sampling*. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria. ISSN 1609-395X.

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, *9*, 501–538. https://doi.org/10.1093/acprof:oso/9780199694587.001.0001

San Martin, E., & Gonzalez, J. (2020). A critical view on the NEAT equating design: Statistical modeling and identifiability problems. *Journal of Educational and Behavioral Statistics*, *47*(4), 406–437. https://doi.org/10.3102/10769986221090609

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling*, *21*(2), 167–180. https://doi.org/10.1080/10705511.2014.882658

Shi, D., Song, H., Liao, X., Terry, R., & Snyder, L. A. (2017). Bayesian SEM for specification search problems in testing factorial invariance. *Multivariate Behavioral Research*, *52*(4), 430–444. https://doi.org/10.1080/00273171.2017.1306432

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

Vats, D., Flegal, J. M., & Jones, G. L. (2015). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, *106*(2), 321–337. https://doi.org/10.1093/biomet/asz002

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (Series B)*, *65*(1), 95–114. 1369-7412/03/65095

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339–361. https://doi.org/10.1177/0146621611405984

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. https://doi.org/10.1214/09-AOS729

Zhang, Y., & Bondell, H. D. (2018). Variable selection via penalized credible regions with Dirichlet-Laplace global–local shrinkage priors. *Bayesian Analysis*, *13*(3), 823–844. https://doi.org/10.1214/17-BA1076

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

(*Appendices follow*)

Appendix A

Implementation of the Shrinkage Priors for the MNLFA

In the following the prior choices as well as the hyperpriors selected for the simulation study will be presented.

Across all implementations, the following priors were chosen (for a single factor model):

$$\nu_{0k} \sim N(0, 1), \quad k = 1, \ldots, 6 \tag{A1}$$

$$\lambda_{0k} \sim N(1, 1), \quad k = 1, \ldots, 6 \tag{A2}$$

$$\eta_i \sim N(0, 1), \quad i = 1, \ldots, N \tag{A3}$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. In the case of continuous items, the prior for the residual variance were specified as

$$\sigma_k^{-2} \sim \text{Ga}(9, 4), \quad k = 1, \ldots, 6 \tag{A4}$$

For the relevant DIF-related parameters let θ_{kj} be the element in $\mathbf{K}$ or $\mathbf{\Omega}_1$, respectively, that represents DIF for the kth indicator variable ($k = 1 \ldots p$) due to the jth covariate ($j = 1 \ldots q$).

Bayesian Lasso Prior

Both adaptive and nonadaptive Laplace priors were included based on Hans (2009), Park and Casella (2008), and Leng et al. (2014):

$$\theta_{kj} \sim \text{dexp}(\sigma_k \phi_{(kj)}), \tag{A5}$$

$$\phi_{(kj)}^{-2} \sim \text{Ga}(9, 4), \tag{A6}$$

where dexp(s) is the double exponential distribution with scale s and where σ_k is specified in Equation A4. This was implemented as

$$\theta_{kj} \sim N(0, \psi_{kj}\sigma_k^2\phi_{(kj)}^2), \tag{A7}$$

$$\psi_{kj} \sim \exp(1/2), \tag{A8}$$

$$\phi_{(kj)}^{-2} \sim \text{Ga}(9, 4), \tag{A9}$$

where exp(r) is the exponential distribution with rate r.

Dirichlet Laplace Prior

Based on Bhattacharya et al. (2015), the prior is specified as

$$\theta_{kj} \sim \text{dexp}(\tau_{kj}\phi_k) \quad \text{implemented as } \theta_{kj} \sim N(0, \psi_{kj}\tau_{kj}^2\phi^2), \tag{A10}$$

$$\psi_{kj} \sim \exp(1/2), \tag{A11}$$

$$\tau_{\text{k1...q}} \sim \text{Dir}(1 \ldots 1), \tag{A12}$$

$$\phi_k^{-2} \sim \text{Ga}(q, 0, 5), \tag{A13}$$

where q is the number of covariates. For the Dirichlet distribution, a diffuse prior was chosen with $a = 1$. This reflects a situation where no information about the number of DIF items is used.

Horseshoe Prior

The horseshoe was implemented with the specification as provided by Carvalho et al. (2010) and Bhadra et al. (2017):

$$\theta_j \sim N(0, \tau_{kj}^2\phi_k^2), \tag{A14}$$

$$\tau_{kj} \sim C^+(0, 1), \tag{A15}$$

$$\phi_k \sim C^+(0, 1), \tag{A16}$$

where $C^+(0, \cdot)$ indicates a half-Cauchy distribution that is truncated at zero.

Spike and Slab Prior

The original spike-and-slab prior is given by Hans (2010)

$$\theta_{kj} \sim (1 - \phi_{kj})\delta_0 + \phi_{kj} \text{ dexp } (\sigma_k\tau_{kj}) \text{ implemented as} \tag{A17}$$

$$\theta_{kj} \sim \phi_{kj}\, N(0, \psi_{kj}\sigma_k^2\tau_{kj}^2), \tag{A18}$$

$$\psi_{kj} \sim \exp(1/2), \tag{A19}$$

$$\phi_{kj} \sim \text{Bern}(\pi_{kj}), \tag{A20}$$

$$\pi_{kj} \sim \text{Beta}(1, 1), \tag{A21}$$

$$\tau_{kj}^{-2} \sim \text{Ga}(9, 4), \tag{A22}$$

and σ_k from Equation A4 above. δ_0 is a point mass at zero.

Here, we used an alternative implementation (Brandt et al., 2018):

$$\theta_{kj} \sim (1 - \phi_{kj})\delta_0 + \phi_{kj} \text{ dexp } (\sigma_k\tau_{kj}) \text{ implemented as,} \tag{A23}$$

$$\theta_{kj} \sim \phi_{kj}\, N(0, \psi_{kj}\sigma_k^2\tau_{kj}^2), \tag{A24}$$

$$\psi_k \sim \exp(1/2), \tag{A25}$$

$$\phi_{kj} \sim \text{Beta}(1, 1), \tag{A26}$$

$$\tau_{kj}^{-2} \sim \text{Ga}(9, 4). \tag{A27}$$

(*Appendices continue*)

The continuous mixture version developed by Ishwaran and Rao (2005) is implemented as

$$\theta_{kj} \sim N(0, \tau_{kj}\nu^2), \tag{A28}$$

$$\nu_j^{-2} \sim \text{Ga}(9, 4), \tag{A29}$$

$$\tau_{kj} \sim (1 - \phi_{kj})\delta_{\nu_0} + \phi_{kj}\delta_1, \tag{A30}$$

$$\phi_{kj} \sim \text{Beta}(1, 1). \tag{A31}$$

The delta spike is set at $\nu_0 = 0.005$, and it is not zero because then the normal distribution is ill-formed with a zero span.

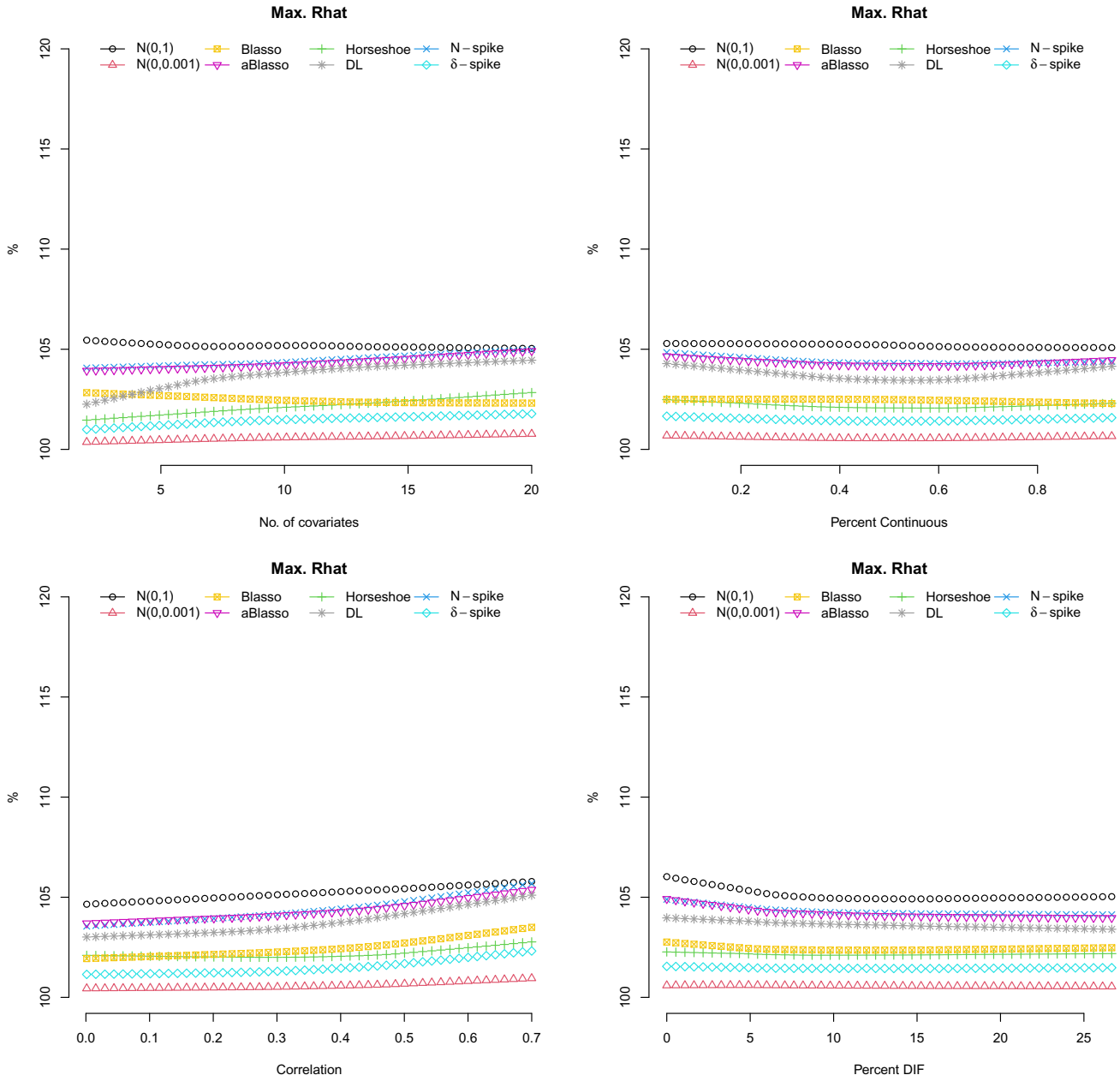
Appendix B

Additional Results for the Simulation Study

See Figure B1.

Figure B1

Relationship Between Convergence Rates and the Four Random Design Factors (Loess Approximation)



Note. See the online article for the color version of this figure.

(*Appendices continue*)

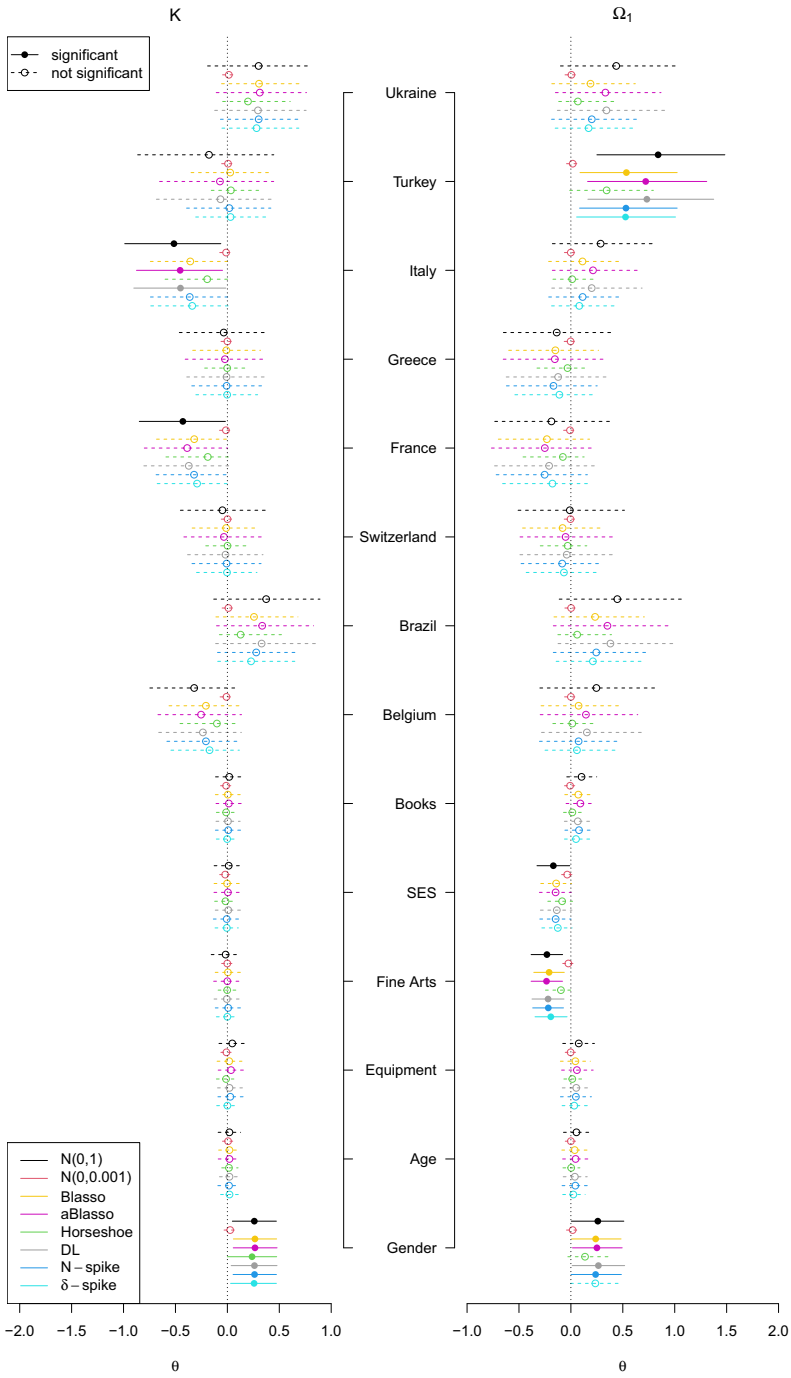
Appendix C

Additional Results From the Empirical Example

See Figures C1, C2, C3, C4, C5, C6, and C7.

Figure C1

*Parameter Estimates and Percentile Intervals for the DIF in Intercepts (**K**) and Factor Loadings (Ω_1) in the Second Item ("I Quickly Read Through the Text Twice") Across the 14 Covariates*

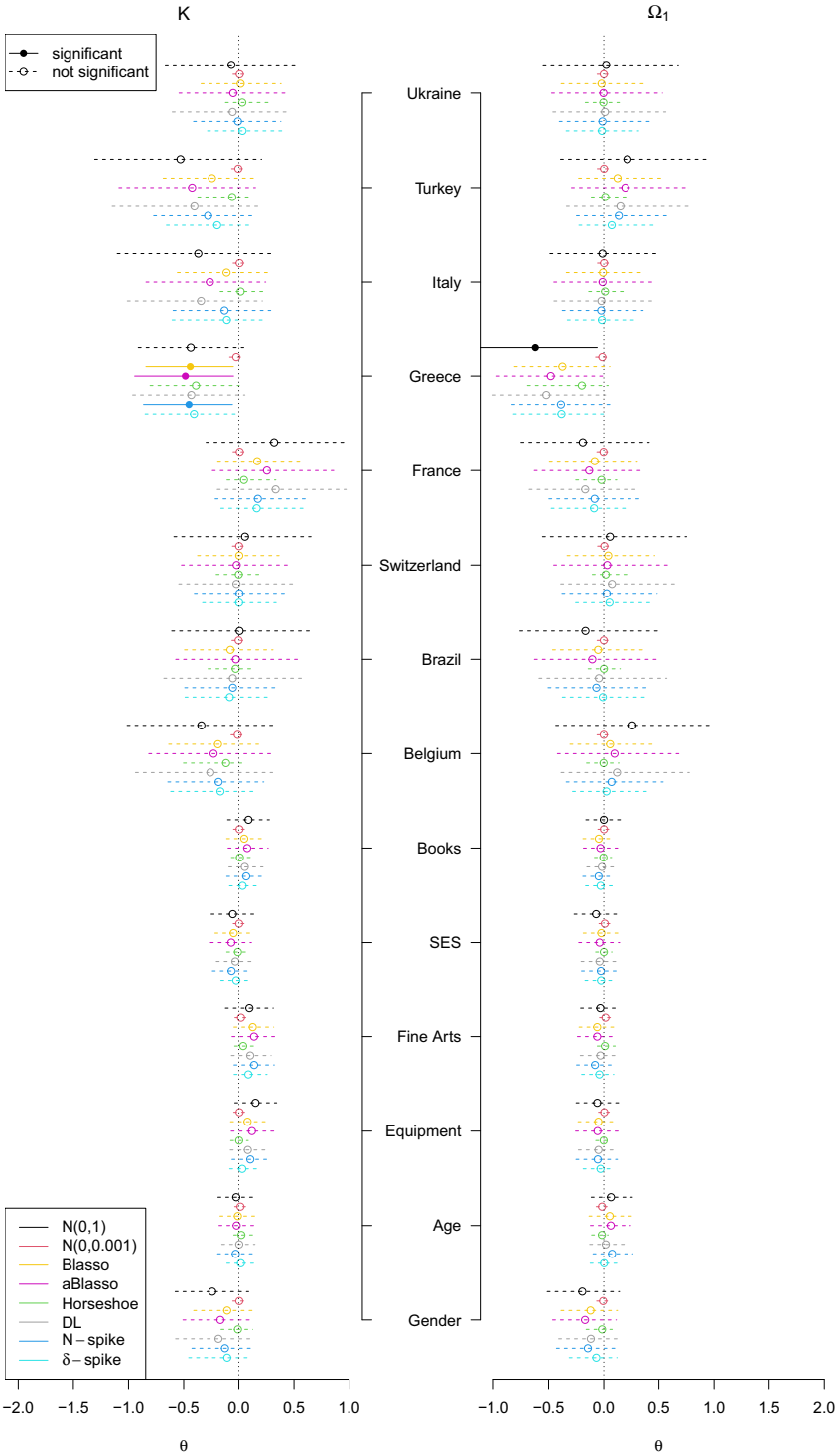


Note. Significant estimates are depicted with solid lines and filled dots. DIF = differential item functioning; SES = socio-economic status. See the online article for the color version of this figure.

(*Appendices continue*)

Figure C2

*Parameter Estimates and Percentile Intervals for the DIF in Intercepts (**K**) and Factor Loadings (Ω_1) in the Third Item ("After Reading the Text, I Discuss its Content With Other People") Across the 14 Covariates*

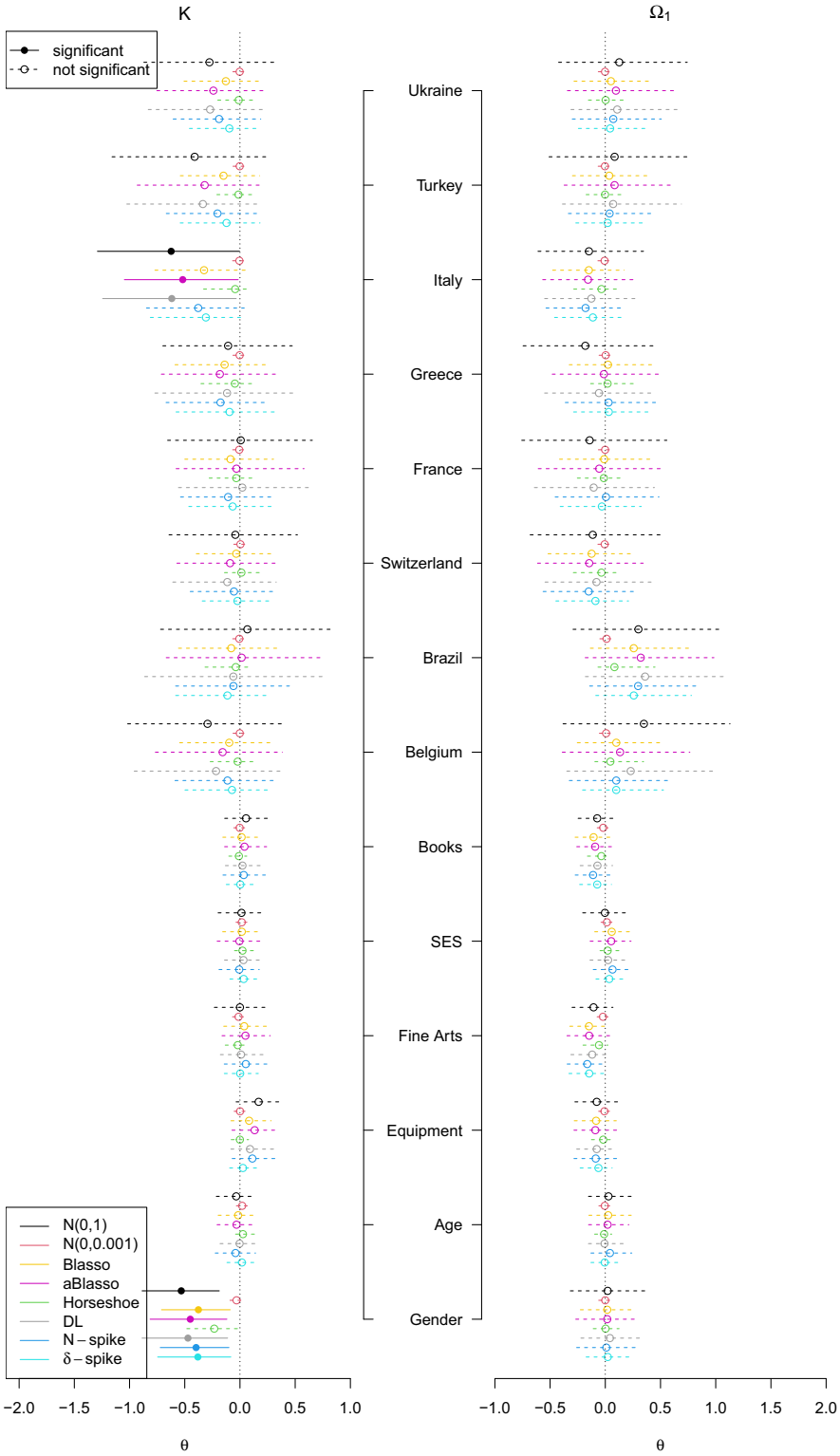


Note. Significant estimates are depicted with solid lines and filled dots. DIF = differential item functioning; SES = socio-economic status. See the online article for the color version of this figure.

(*Appendices continue*)

Figure C3

*Parameter Estimates and Percentile Intervals for the DIF in Intercepts (**K**) and Factor Loadings (Ω_1) in the Fourth Item ("I Underline Important Parts of the Text") Across the 14 Covariates*

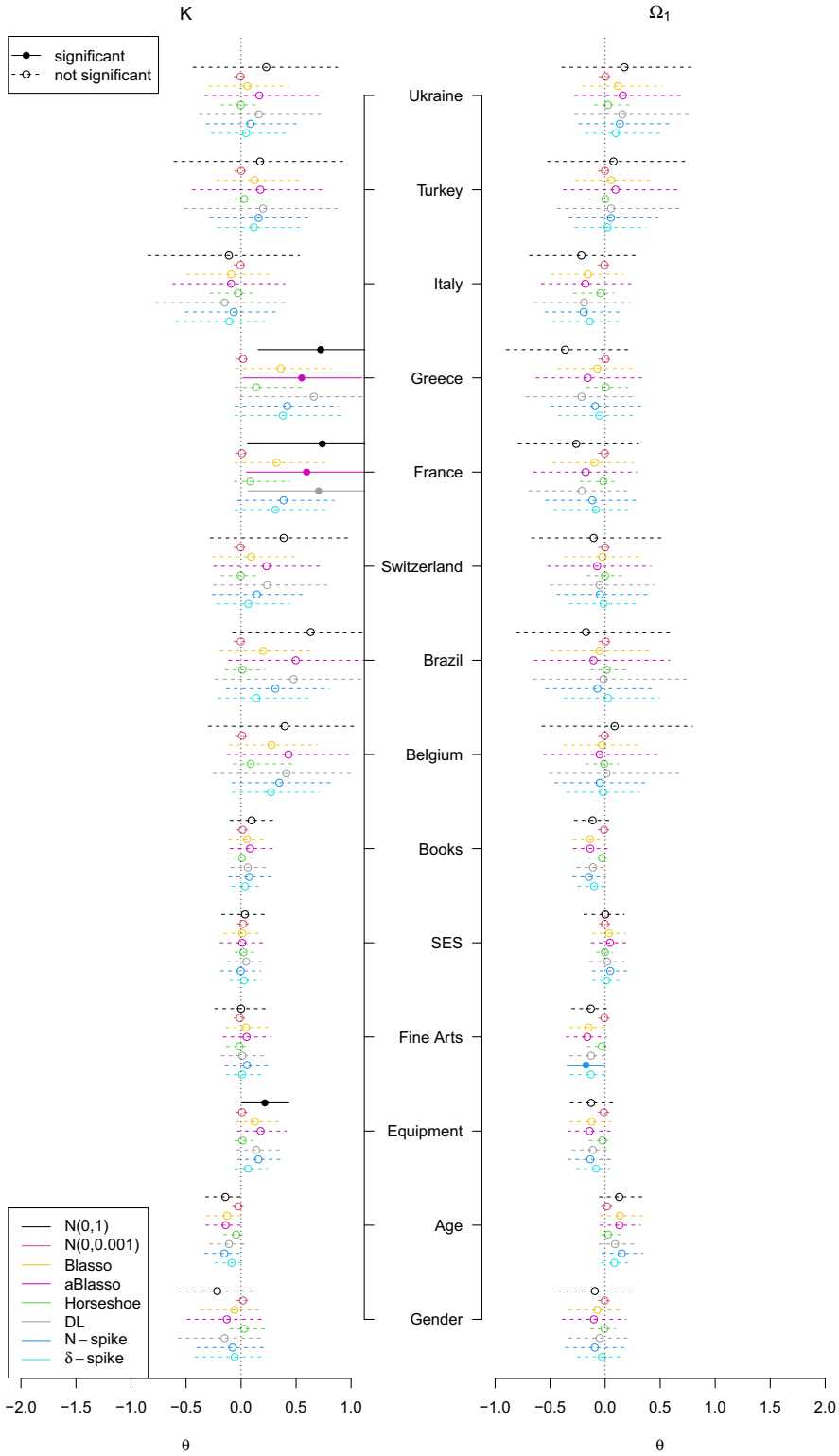


Note. Significant estimates are depicted with solid lines and filled dots. DIF = differential item functioning; SES = socio-economic status. See the online article for the color version of this figure.

(*Appendices continue*)

Figure C4

*Parameter Estimates and Percentile Intervals for the DIF in Intercepts (**K**) and Factor Loadings (Ω_1) in the Fifth Item ("I Summarise the Text in My Own Words") Across the 14 Covariates*

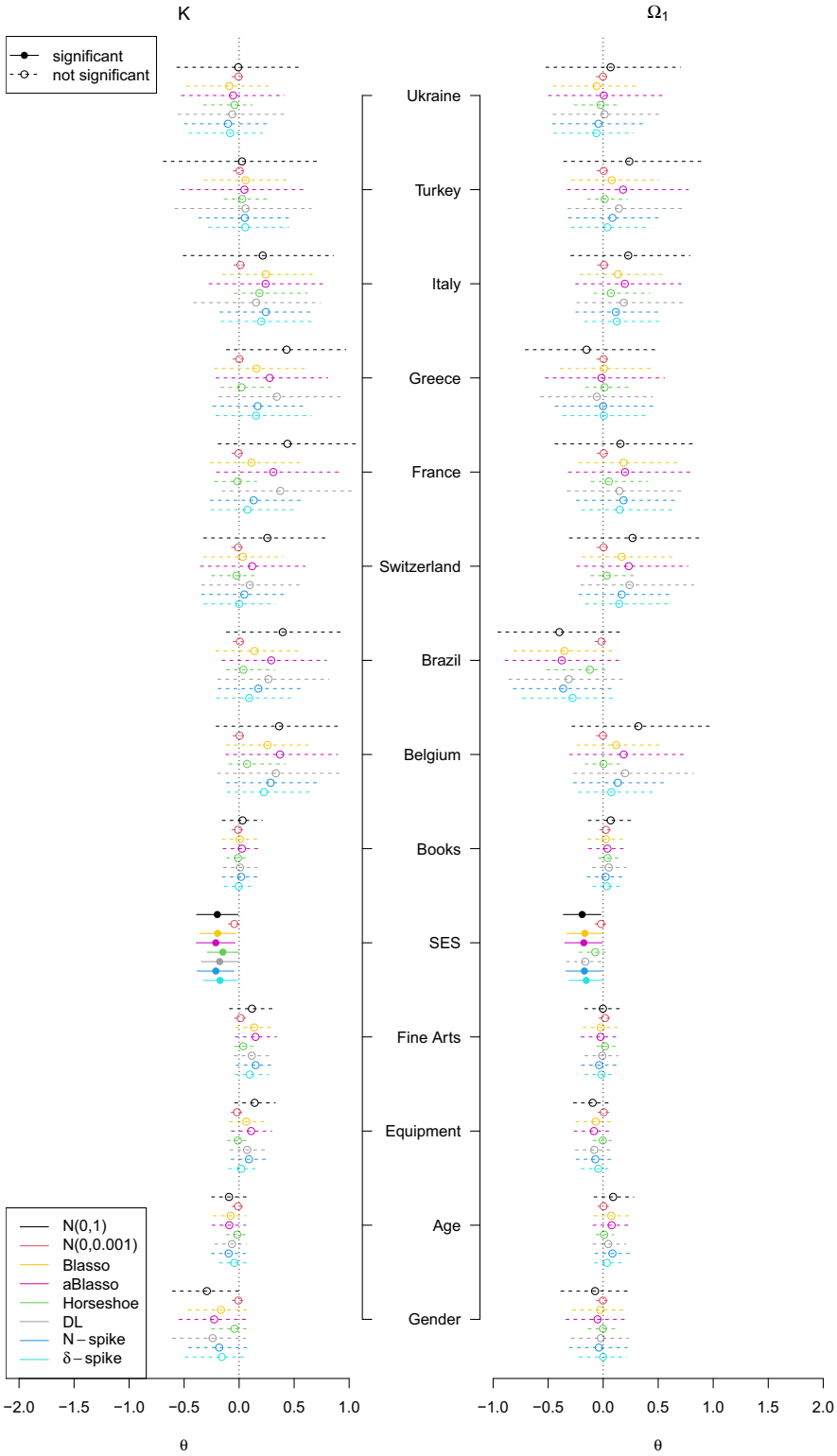


Note. Significant estimates are depicted with solid lines and filled dots. DIF = differential item functioning; SES = socio-economic status. See the online article for the color version of this figure.

(*Appendices continue*)

Figure C5

*Parameter Estimates and Percentile Intervals for the DIF in Intercepts (**K**) and Factor Loadings (Ω_1) in the Sixth Item ("I Read the Text Aloud to Another Person") Across the 14 Covariates*

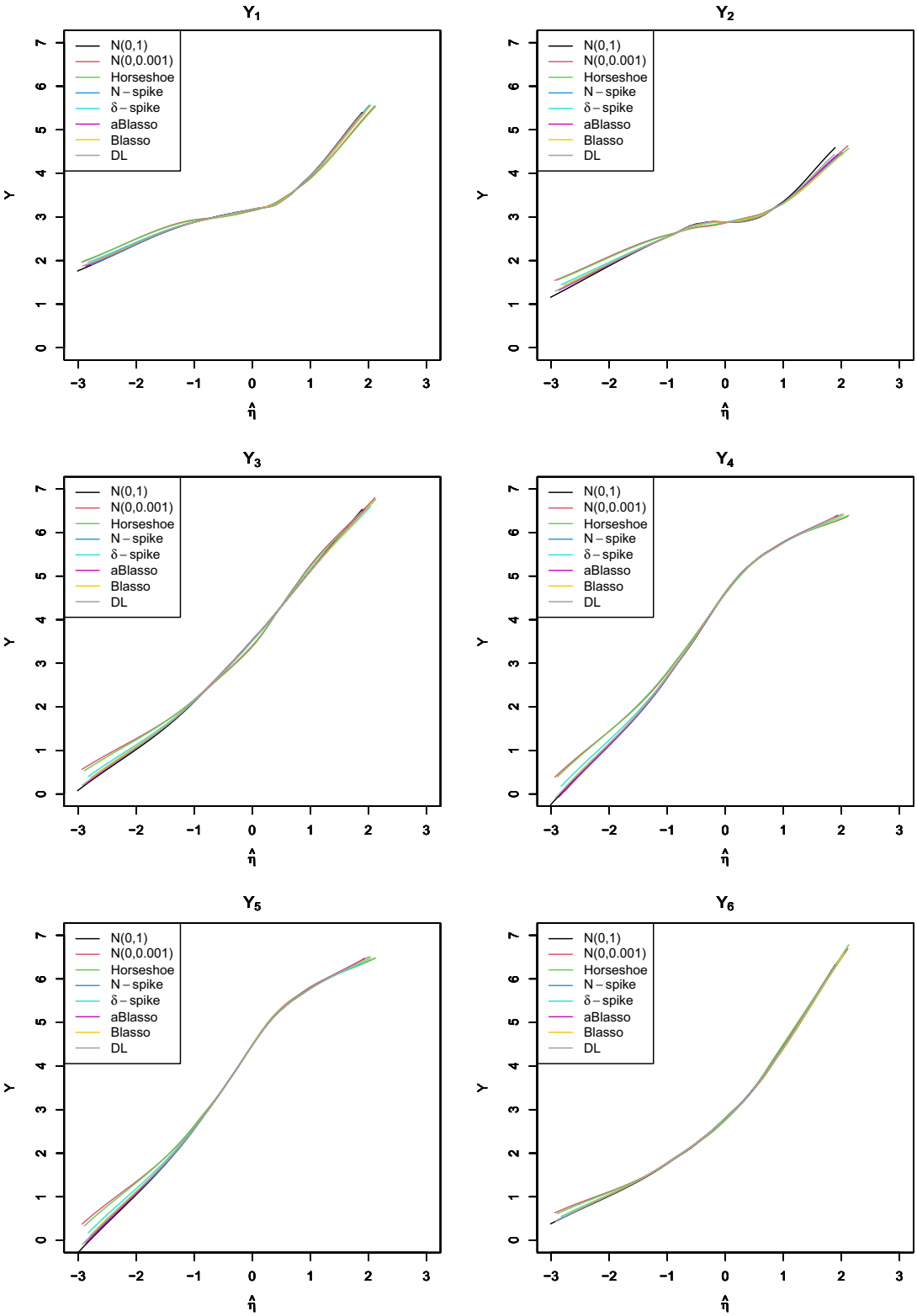


Note. Significant estimates are depicted with solid lines and filled dots. DIF = differential item functioning; SES = socio-economic status. See the online article for the color version of this figure.

(*Appendices continue*)

Figure C6

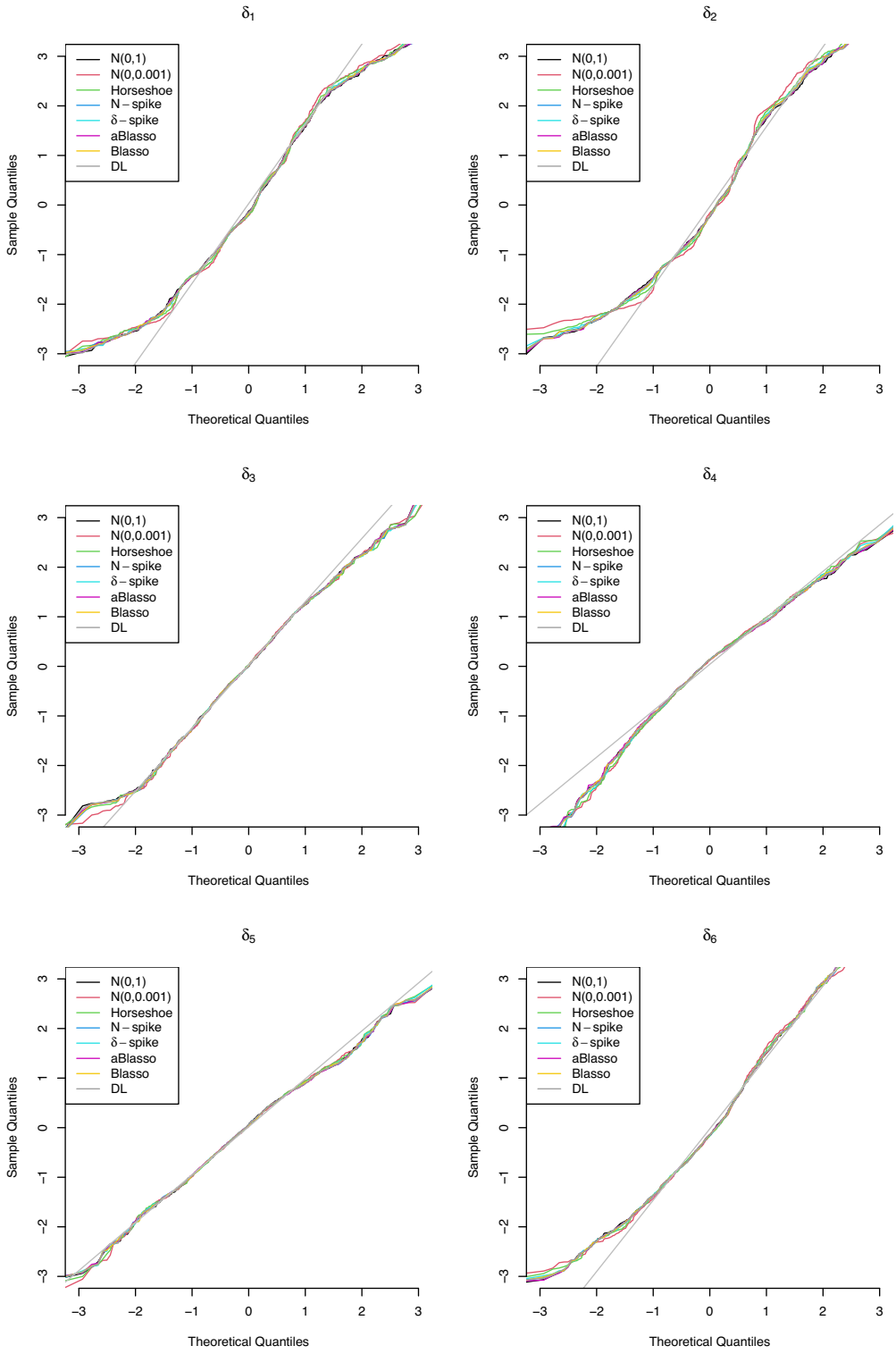
Partial Plots for Each of the Six Items



Note. For the factor scores $\hat{\eta}$, we partialled all other covariates out. The lines indicate loess approximations of the functional relationships between the variables. See the online article for the color version of this figure.

Figure C7

QQ Plots for the Residuals of Each of the Six Items



Note. See the online article for the color version of this figure.

Received March 22, 2022
Revision received October 29, 2022
Accepted November 07, 2022 ■