Psychometrika (2026), **91**, 1048–1068
doi:10.1017/psy.2026.10101

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Bayesian Covariance Modeling of Differential Item Functioning

Jean-Paul Fox

Behavioural, Management and Social Sciences, University of Twente, Netherlands
Email: j.p.fox@utwente.nl

(Received 23 April 2025; revised 11 February 2026; accepted 23 February 2026)

Abstract

In a Bayesian modeling approach for differential item functioning (DIF), the dependence structure of (non-)uniform DIF is represented by a structured covariance matrix. Item–group interactions (uniform DIF) but also item-specific person–group interactions (non-uniform DIF) are represented by additional correlations in (latent) item responses. DIF in discriminations and difficulties is modeled simultaneously across items and multiple groups in the covariance matrix, making it possible to examine (non-)uniform DIF without needing anchor item(s) or multiple-step procedures. The modeling framework is very efficient, avoids the computation of any interaction parameter, and requires only a single covariance parameter for the DIF assessment of an item parameter for any number of groups. This supports a simultaneous non-uniform DIF analysis of all items, even for small sample sizes. The proposed DIF procedure is applied to PISA data, where the advantages of the method are illustrated and compared to a multiple-group IRT analysis.

Keywords: Bayes factor; Bayesian covariance structure; IRT; (non-)uniform DIF

1. Introduction

Measurement bias in the form of differential item functioning (DIF) has received considerable attention in the literature of educational measurement. An item is known to exhibit DIF if the probability of success given ability differs from group to group. Models for DIF detection—also known as measurement invariance testing—have been developed in which an item–group interaction effect represents uniform DIF for an item. The logistic mixed-effects model have been defined for this purpose (Hartig et al., 2020; Van den Noortgate & De Boeck, 2005). Another popular model for DIF analysis is the multiple indicators, multiple causes (MIMIC) confirmatory factor model. In the MIMIC model, the measurement component contains an item-specific group effect, which represents the uniform DIF effect (Chen et al., 2023; Muthèn, 1985). In the multiple-group item response theory (MG-IRT) model group-specific IRT functions can be estimated, which can be compared to the common (invariant) item response functions across groups (Bock & Zimowski, 1997; Köhler et al., 2024).

Methods based on item–group interaction effects (uniform DIF) have the disadvantage that the interactions need to be estimated and assessed for significance. This places restrictions on the required sample size for each group to estimate the interaction effect with sufficient precision. For more than two groups, the assessment of DIF becomes a multiple hypothesis testing problem and requires

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (https://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

pairwise-group comparisons. Furthermore, for non-uniform DIF, the estimation and assessment of the interaction between the level of ability and group membership for an item have proven to be challenging. There are only a few studies reporting about performances of methods for non-uniform DIF detection (Narayanon & Swaminathan, 1996).

A concurrent (non-)uniform DIF analysis of all test items is computationally very challenging due to the large number of interaction effects and the required large sample sizes. Recently, regularized DIF methods have been developed using a lasso-type procedure for model selection and parameter estimation (e.g., Belzak & Bauer, 2020; Schauberger & Mair, 2020), which allows the explicit parameterization of all possible DIF effects for all items with the restriction that no more than 50% of the items show DIF. The regularized DIF (lasso) method supports the identification of DIF only through estimation, where non-zero estimated effects are considered DIF. However, the procedure does not provide constructs like *P*-values and confidence intervals of the lasso (DIF) estimates, making it impossible to assess the significance of the estimated DIF effects. Belzak and Bauer (2020) recommended re-fitting the final model to compute *P*-values and confidence intervals of the DIF parameters given selected anchor items obtained through regularization. However, they also reported that this leads to naive intervals and *P*-values—ignoring the impact of the data-dependent model selection on statistical inference—and intervals are too short and will show under coverage. The uncertainty regarding the selection of the DIF items needs to be integrated into the statistical inference to guarantee its validity. However, performing inference after data-driven model selection is a very complex task but inherently connected to lasso-type procedures (Zhang et al., 2022).

Valid and consistent methods based on model comparisons require a baseline model, which requires known anchor items. Constrained (nested) versions of the baseline model can be tested to detect (non-uniform) invariant items (Chalmers, 2012). To test a set of items for DIF, a sequential test procedure is required, for which it is not exactly clear in what order to test the items. When interest is in DIF in discriminations and difficulties, the number of restricted models to examine can increase quickly, since constrained versions of the baseline model are needed for both parameters. Furthermore, sequentially or iteratively testing for DIF leads to a multiple hypothesis testing problem, which requires a correction to control the family-wise error rate.

1.1. The disadvantages of random item parameters

To enable a simultaneous DIF analysis of all items, and to avoid a model-selection procedure, Bayesian IRT models for DIF have been developed (e.g., De Jong et al., 2007; Verhagen & Fox, 2013). They are constructed from a random item parameter model, also known as a random item effects model, in which random item parameters describe differences in item parameters across groups. The model defines a common measurement scale across groups by assuming group-specific (random) item parameters, which are assumed to be normally distributed around invariant item parameters. The random item parameter variance is used to assess the degree of measurement variance. When the variance is (approximately) zero, it represents no DIF (measurement invariance), and if greater than zero, it represents DIF (measurement variance).

The random item parameter approach has several disadvantages. The method is not suitable for non-sampled groups for which the normality assumption of the random item parameters does not apply, or a few sampled groups for which the random item parameter variance cannot be accurately estimated. Another disadvantage is that DIF is represented by a random item parameter variance of zero, which is a boundary value on the parameter space. This complicates the specification of an objective prior for the variance parameter but also complicates assessing data evidence in favor of DIF under the random item parameter model. Furthermore, the random item parameter model, similar to DIF interaction models, places restrictions on the required group sizes to estimate random item effects with sufficient precision. This is particularly a problem for examining DIF in discrimination parameters, which relies on accurate group-specific discrimination estimates.

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

1.2. From random item parameters to a structured covariance matrix

To avoid the disadvantages of the random item parameter model, Fox et al. (2017) and Fox et al. (2020) proposed a Bayesian covariance structure model for DIF testing (BCSM-MI). In their modeling approach, dependence among item responses caused by DIF is not modeled by random item parameters but through a structured covariance matrix. When assuming DIF, the IRT's local independence assumption does not hold, which means that the item response data are not independently distributed given person and invariant item parameters. The item response data are assumed to be correlated within each group, caused by an item–group interaction. The correlation is zero when measurement invariance holds, and it is greater than zero when measurement invariance is violated.

The BCSM-MI has the advantage that it does not include any group-specific item parameters, and subsequently does not make any assumptions about the distribution of the group-specific item parameters. This makes the BCSM-MI suitable for assessing DIF for two or more groups and for randomly and non-randomly selected groups. Furthermore, a correlation of zero (representing no DIF) is not a boundary value on the parameter space. This makes it possible to define a noninformative prior about the DIF assumption and to assess data evidence in favor of DIF under the BCSM-MI. Furthermore, the assumption of uniform DIF (item–group interaction) is represented by a single covariance parameter, but also non-uniform DIF (item-specific person–group interaction) is represented by a single covariance parameter. Without requiring any item–group interaction effects, the assessment of DIF in discriminations and difficulties can be executed with more similar levels of confidence.

In this study, the BCSM-MI for uniform DIF is generalized to assess non-uniform DIF for unbalanced groups. DIF is manifested through differences in difficulty or discrimination or both across groups. Although Fox et al. (2020) introduced a structured covariance matrix representing the assumption of uniform and non-uniform DIF for a balanced design, their developed Markov chain Monte Carlo (MCMC) algorithm and simulation results were restricted to uniform DIF. Their MCMC algorithm is extended to include the estimation of (item-specific) covariance parameters—a dependence structure implied by item-specific person–group interactions—while also accounting for unequal group sizes. The conditional distribution of each covariance parameter is identified as a product of shifted-inverse gamma distributions (Fox et al., 2017), which is shown through eigenvector decompositions. Lower bound restrictions for the covariance parameters, ensuring a positive-definite covariance matrix, are integrated in the shifted-inverse gamma priors. This facilitates efficient posterior sampling of covariance parameters as part of the MCMC algorithm.

Next, the BCSM-MI is introduced, in which a structured covariance matrix represents the assumption of uniform and non-uniform DIF. Subsequently, a short comparison is given with some conventional DIF tests, and shortcomings in relation to the proposed Bayesian DIF tests are discussed. In Section 4, the proposed inference method is described, including the derivation of the posterior distribution of the covariance parameters. In Section 5, a Bayes factor (BF) test is proposed for assessing hypotheses concerning non-uniform DIF. In Section 6, the performance of the BCSM-MI for the assessment of non-uniform DIF is shown through simulation studies. An illustration of the method is given using a PISA study and results are compared to an MG-IRT analysis. Finally, in Section 8, a discussion is given of possible extensions of the method and of the used covariance priors.

2. The BCSM-MI method

The BCSM-MI is introduced by considering the random item parameter model for binary item responses (Fox, 2010, Verhagen & Fox, 2013). The random item parameters, denoted as a_{kj} and b_{kj}, are assumed to be normally distributed with mean a_k and b_k and variance σ_{a_k} and σ_{b_k}, respectively. A latent response variable is used to specify the random item parameter model with Z_{ijk} a normally distributed latent response of person i in group j for item k. Then, the random item effects model is represented by

$$\begin{aligned} Z_{ijk} &= a_{kj}\theta_{ij} - b_{kj} + e_{ijk}, \\ &= a_k\theta_{ij} - b_k + \varepsilon_{a_{kj}}\theta_{ij} + \varepsilon_{b_{kj}} + e_{ijk}, \end{aligned} \tag{1}$$

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

where $\varepsilon_{a_{kj}}\theta_{ij}$ and $\varepsilon_{b_{kj}}$ represents an item's person–group interaction (DIF in discrimination) and an item–group interaction (DIF in difficulty) in group j, respectively, and assumed to be normally distributed with mean zero and variance σ_{a_k} and σ_{b_k}, respectively. The random measurement error e_{ijk} is assumed to be standard normally distributed. Following Fox et al. (2020), the BCSM-MI is derived by integrating out the random item parameters. Without conditioning on the group-specific item parameters, they show that the covariance among two latent responses to item k in group j is given by

$$Cov\left(Z_{ijk}, Z_{ljk} \mid \theta_{ij}, a_k, b_k\right) = \sigma_{a_k}\theta_{ij}\theta_{lj} + \sigma_{b_k}. \tag{2}$$

Given this result, the BCSM-MI for responses to item k in group j, including the dependence structure of the DIF effects, can be represented by

$$\begin{aligned} \mathbf{Z}_{jk} &\sim \mathcal{N}\left(a_k\boldsymbol{\theta}_j - b_k, \boldsymbol{\Sigma}_{jk}\right) \\ \boldsymbol{\Sigma}_{jk} &= \mathbf{I}_{n_j} + \sigma_{b_k}\mathbf{J}_{n_j} + \sigma_{a_k}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t, \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_j \sim \mathcal{N}\left(\mu_{\theta_j}, \sigma_{\theta_j}\mathbf{I}_{n_j}\right)$ and $\mathbf{Z}_{jk} \in \boldsymbol{\Omega}(\mathbf{z}_{jk})$ with $\boldsymbol{\Omega}(\mathbf{z}_{jk}) = \{\mathbf{z}_{jk}, z_{ijk} \leq 0 \text{ if } y_{ijk} = 0, z_{ijk} > 0 \text{ if } y_{ijk} = 1\}$ representing the allowable region for the latent responses. The $\mathbf{J}_{n_j}$ is a matrix of dimension n_j with all elements equal to one and $\mathbf{I}_{n_j}$ is the identity matrix.

To illustrate the modeling of DIF through the structured covariance matrix, the conditional expectation of Z_{ijk} is derived under the BCSM-MI given item and person parameters and remaining latent responses in group j, denoted as $\mathbf{Z}_{(-i)jk}$. When $\sigma_{b_k} = 0$ representing no DIF in difficulty parameter b_k, it is shown that this conditional expectation equals the mean term $\mu_{ijk} = a_k\theta_{ij} - b_k$ plus an item–person–group interaction, $\theta_{ij}\varepsilon_{a_{kj}}$, according to Equation (1). From Equation (1), with $\varepsilon_{b_{kj}} = 0$, it follows that the ordinary least squares estimator of $\varepsilon_{a_{kj}}$ can be expressed as

$$\hat{\varepsilon}_{a_{kj}} = \frac{\sum_{s\neq i}\theta_{sj}\left(Z_{sjk} - \mu_{sjk}\right)}{\sum_{s\neq i}\theta_{sj}^2}, \tag{4}$$

where observation Z_{ijk} is excluded in the estimator for the $\varepsilon_{a_{kj}}$.

According to the BCSM-MI, the conditional expectation of Z_{ijk} given $\mathbf{Z}_{(-i)jk}$ can be expressed as (see Appendix A)

$$\begin{aligned} E\left(Z_{ijk} \mid \mathbf{Z}_{(-i)jk}\right) &= a_k\theta_{ij} - b_k + \theta_{ij}\frac{\sum_{s\neq i}\theta_{sj}\left(Z_{sjk} - \mu_{sjk}\right)}{1/\sigma_{a_k} + \sum_{s\neq i}\theta_{sj}^2} \\ &= a_k\theta_{ij} - b_k + \theta_{ij}\left(\frac{\sum_{s\neq i}\theta_{sj}^2}{1/\sigma_{a_k} + \sum_{s\neq i}\theta_{sj}^2}\right)\hat{\varepsilon}_{a_{kj}}, \end{aligned} \tag{5}$$

where, in the last term, a Bayes estimator of the $\varepsilon_{a_{kj}}$ can be recognized, in which a positive σ_{a_k} arranges shrinkage of the effect toward zero. As a conclusion, under the BCSM-MI, the expected latent response Z_{ijk} depends on the item–person–group interaction, which represents the contribution of DIF in discrimination parameter a_k. When $\sigma_{a_k} = 0$, the covariance support for this interaction effect is not present in the structured covariance matrix in Equation (3). Subsequently, the expression in Equation (5) is only defined for $\sigma_{a_k} \neq 0$.

The structured covariance matrix needs to be positive definite, which restricts the parameter space of the covariance parameters. When $\sigma_{b_k} = 0$, the determinant of the covariance matrix needs to be positive and implies a restriction on the σ_{a_k}. It follows that

$$\begin{aligned} det\left(\mathbf{I}_{n_j} + \sigma_{a_k}\boldsymbol{\theta}_j\boldsymbol{\theta}_j\right) &= det\left(\mathbf{I}_{n_j}\right)\left(1 + \sigma_{a_k}\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j\right) > 0 \\ &= 1 + \sigma_{a_k}\sum_i\theta_{ij}^2 > 0. \end{aligned} \tag{6}$$

Subsequently, the covariance matrix is positive definite if $\sigma_{a_k} > -1/\sum_i\theta_{ij}^2$. The parameter space of the covariance parameter σ_{a_k} represents three different types of non-uniform DIF. The common

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

non-uniform DIF in discrimination is represented by $\sigma_{a_k} > 0$, which states a non-uniform difference in item characteristic curves across groups. There is no DIF in discrimination if $\sigma_{a_k} = 0$. When $\sigma_{a_k} \in \left(-1/\sum_i \theta_{ij}^2, 0\right)$, there is DIF in discrimination within each group j. This means that for persons in group j, the item discriminates differently and this is referred to as within-group non-uniform DIF. Within each group, the item discriminates more highly to some than other group members, and this is manifested as a negative correlation among the (latent) item responses. Within-group DIF is beyond the scope of this research. Nielsen et al. (2021) and Fox and Smink (2023) discuss the modeling of negative associations among clustered measurements and the interpretation of negative clustering effects.

3. Related methods on detecting DIF

The psychometric literature on DIF is huge and shows intensive research in assessing the fairness of measurement instruments. The IRT-based methods condition on the latent variable and are originally designed for a two-group comparison (Camilli & Shepard, 1994; Millsap, 2011). A likelihood-ratio (LR) test can be applied to test for differences in item parameters across groups (Thissen et al., 1988). These methods require prior knowledge about anchor items assumed to be free of DIF. Results from a stepwise procedure to identify anchor items can be sensitive to the initial choice of anchors. Uncertainty in the model selection process can lead to an incorrect selection of items free of DIF. Due to the stepwise procedure of model selection, the validity of the statistical inference on the DIF parameters is not guaranteed. Disadvantages of DIF methods based on pre-knowledge of anchoring items have been discussed by Chen et al. (2023).

Recent extensions in IRT-based DIF modeling consider additional group-specific parameters for differences across groups in item discrimination and difficulty for several items. Although an IRT-based LR test can be used to identify variations in item discriminations across groups, due to the increase in the number of model parameters maximum likelihood estimation is more likely to fail. Furthermore, the LR method seriously increases the required sample size in each group for accurate parameter estimation.

The performance of DIF detection methods have not been extensively studied, mainly due to the computational burden. Finch (2005) provides an overview of the performance of several popular methods (e.g., MIMIC, Mantel–Haenszel, and LR test), which require anchor items and they are meant for a two-group comparison. It is shown that the performances depend on the number of anchor items, where more anchors increased the power of DIF detection. Anchors containing DIF negatively affect the performances. Furthermore, the Type-1 error was beyond the nominal level especially for relatively short tests (around 20 items). For longer tests, the MIMIC model showed only a small inflation in the Type-I error. Shih and Wang (2009) examined the performance of the MIMIC model for DIF detection and showed acceptable (average) Type-I error rates. However, for a group size of 500 and 1,000, the mean power (true positive rate) was around 50%–60% and 80%–90%, respectively, where the difference in difficulty parameters between the reference and focal groups was set at 0.4. It is well known that conditional DIF methods, in which DIF effects are estimated and assessed for significance, require moderate to large sample sizes to perform well (Fox et al., 2020).

Due to page limit, a further discussion of more recently developed DIF methods (e.g., regularized DIF and model-based recursive partitioning) is given in Section S1 of the Supplementary Material. Subsequently, in Section S2 of the Supplementary Material, a comparison in performance between regularized DIF and the BCSM-MI method is given through a simulation study.

The BCSM-MI has several advantages over other DIF methods. For each item parameter, DIF can be assessed for any number of groups with a single covariance parameter. The BCSM-MI method avoids the model selection problem to detect anchor items by modeling simultaneously DIF in all items. The item DIF is simultaneously assessed for significance, which leads to a valid inference procedure. The proposed BF test for a covariance parameter is an omnibus test for (non-)uniform DIF across all groups thereby avoiding multiple hypothesis testing problem. The method does not rely on any anchor items to measure all persons on the same scale. The DIF in each item is modeled through the covariance matrix,

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

which adjusts the person and item parameter estimates but this does not lead to changes in the scale of the person parameters. Therefore, the BCSM-MI can be identified by restricting the mean and variance of the scale of the person parameters (possibly of a reference group), independent of the number of items showing DIF.

4. MCMC estimation of BCSM-MI

The MCMC algorithm for the estimation of the BCSM-MI parameters is described in Appendix C. The sampling of the covariance parameters is done through slice sampling and discussed in more detail in this section. The other model parameters are sampled directly from their conditional distributions given augmented item response data. A Metropolis–Hastings step is defined for the sampling of person parameters, when negative covariance values are sampled (see Appendix D).

The BCSM-MI is identified in the same way as an MG-IRT model. The mean and variance of the person parameters of the reference group are restricted to 0 and 1, respectively. It is also possible to place restrictions on the item difficulty and discrimination parameters to identify the latent scale.

4.1. Posterior distribution (DIF) covariance parameters

To obtain an analytical expression for the posterior distribution of each covariance parameter, a novel method is applied based on eigenvector transformations of the latent item response data. This method generalizes the procedure of Fox (2024), in which one covariance parameter is considered. The method is based on recognizing that the covariance structure of the BCSM-MI in Equation (3) consists of two rank-one components to describe dependence among item responses due to DIF. The two non-zero eigenvalues can be expressed as a function of the covariance parameters. The transformation of the (latent) item response data with the eigenvectors is considered, which has a bivariate normal distribution with a covariance matrix containing the two non-zero eigenvalues. Another set of eigenvectors can be computed which diagonalize this covariance matrix with respect to one of the covariance parameters. Then, an analytical expression of each covariance parameter is obtained by considering the distribution of the latent item response data transformed by both eigenvectors.

By definition, the transformation of the (latent) response data by eigenvectors is sufficient for the covariance parameters. Furthermore, any correlation between the DIF in difficulty and discrimination will not bias the estimation results. The second eigenvector decomposition is an orthogonal transformation with respect to the other covariance parameter thereby properly avoiding the influence of any correlation between the two covariance parameters.

4.1.1. Eigenvector transformations

This procedure starts with the computation of the eigenvectors with non-zero eigenvalues for the two rank one components in the covariance matrix. The following definitions follow directly from theory of a rank-1 matrix. Let $\mathbf{v}_{a_{jk}} = \boldsymbol{\theta}_j / \|\boldsymbol{\theta}_j\|$ be the eigenvector which diagonalize the covariance matrix component $\sigma_{a_k}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t$ with eigenvalue $\sigma_{a_k}\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j$, where $\|\boldsymbol{\theta}_j\|$ represents the Frobenius norm of $\boldsymbol{\theta}_j$. The eigenvector $\mathbf{v}_{b_{jk}} = \mathbf{1}_{n_j}/\|\mathbf{1}_{n_j}\|$ diagonalizes the covariance matrix component $\sigma_{b_k}\mathbf{J}_{n_j}$ with eigenvalue $\sigma_{b_k}n_j$. The eigenvectors are stored in matrix $\mathbf{v}_{jk} = (\mathbf{v}_{a_{jk}}, \mathbf{v}_{b_{jk}})$. Then, $\tilde{\mathbf{z}}_{jk} = \mathbf{v}_{jk}^t(\mathbf{z}_{jk} - \boldsymbol{\mu}_{jk})$ has a bivariate normal distribution with mean zero and covariance matrix $\mathbf{v}_{jk}^t\boldsymbol{\Sigma}_{jk}\mathbf{v}_{jk}$.

Consider the projection matrix $\mathcal{P}_a$ and $\mathcal{P}_b$ orthogonal to the column space of $\mathbf{v}_{jk}^t\mathbf{1}_{n_j}$ and $\mathbf{v}_{jk}^t\boldsymbol{\theta}_j$, respectively, which are given by

$$\mathcal{P}_a = \mathbf{I}_2 - \left(\mathbf{v}_{jk}^t\mathbf{1}_{n_j}\right)\left(\left(\mathbf{v}_{jk}^t\mathbf{1}_{n_j}\right)\left(\mathbf{v}_{jk}^t\mathbf{1}_{n_j}\right)^t\right)^{-1}\left(\mathbf{v}_{jk}^t\mathbf{1}_{n_j}\right)^t,$$

$$\mathcal{P}_b = \mathbf{I}_2 - \left(\mathbf{v}_{jk}^t\boldsymbol{\theta}_j\right)\left(\left(\mathbf{v}_{jk}^t\boldsymbol{\theta}_j\right)\left(\mathbf{v}_{jk}^t\boldsymbol{\theta}_j\right)^t\right)^{-1}\left(\mathbf{v}_{jk}^t\boldsymbol{\theta}_j\right)^t,$$

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

respectively. The projection matrices $\mathcal{P}_a$ and $\mathcal{P}_b$ are each of rank one. Let $\mathbf{w}_{a_{jk}}$ and $\mathbf{w}_{b_{jk}}$ be the eigenvector of projection matrices $\mathcal{P}_a$ and $\mathcal{P}_b$, respectively, each corresponding to the one non-zero eigenvalue. Then, by definition, $\mathbf{w}_{a_{jk}}$ and $\mathbf{w}_{b_{jk}}$ diagonalize the covariance matrix $\mathbf{v}_{jk}^t\boldsymbol{\Sigma}_{jk}\mathbf{v}_{jk}$ with respect to σ_{a_k} and σ_{b_k}, respectively. For instance, eigenvectors $\mathbf{w}_{a_{jk}}$ and $\mathbf{v}_{jk}$ diagonalize covariance matrix $\boldsymbol{\Sigma}_{jk}$ with respect to σ_{a_k}:

$$\begin{aligned}\mathbf{w}_{a_{jk}}^t\left(\mathbf{v}_{jk}^t\boldsymbol{\Sigma}_{jk}\mathbf{v}_{jk}\right)\mathbf{w}_{a_{jk}} &= \mathbf{w}_{a_{jk}}^t\left(\mathbf{v}_{jk}^t\mathbf{I}_{n_j}\mathbf{v}_{jk}+\sigma_{b_k}\mathbf{v}_{jk}^t\mathbf{J}_{n_j}\mathbf{v}_{jk}+\sigma_{a_k}\mathbf{v}_{jk}^t\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t\mathbf{v}_{jk}\right)\mathbf{w}_{a_{jk}}\\ &= \mathbf{w}_{a_{jk}}^t\mathbf{v}_{jk}^t\mathbf{I}_{n_j}\mathbf{v}_{jk}\mathbf{w}_{a_{jk}}+\sigma_{a_k}\mathbf{w}_{a_{jk}}^t\mathbf{v}_{jk}^t\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t\mathbf{v}_{jk}\mathbf{w}_{a_{jk}},\\ &= \tilde{\mathbf{w}}_{a_{jk}}^t\tilde{\mathbf{w}}_{a_{jk}}+\sigma_{a_k}\left(\tilde{\mathbf{w}}_{a_{jk}}^t\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t\tilde{\mathbf{w}}_{a_{jk}}\right),\end{aligned}$$

where $\tilde{\mathbf{w}}_{a_{jk}}^t = \mathbf{w}_{a_{jk}}^t\mathbf{v}_{jk}^t$ and $\tilde{\mathbf{w}}_{a_{jk}}^t\mathbf{1}_{n_j} = \mathbf{0}$, since $\mathcal{P}_a$ is orthogonal to the column space of $\mathbf{v}_{jk}^t\mathbf{1}_{n_j}$.

4.1.2. Posterior distribution of σ_{a_k}

Consider the distribution of $\tilde{\tilde{z}}_{a_{jk}} = \tilde{\mathbf{w}}_{a_{jk}}^t\tilde{\mathbf{z}}_{jk}$. Let $s_{a_{jk}} = \tilde{\mathbf{w}}_{a_{jk}}^t\tilde{\mathbf{w}}_{a_{jk}}$ and $\lambda_{a_{jk}} = \tilde{\mathbf{w}}_{a_{jk}}^t\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t\tilde{\mathbf{w}}_{a_{jk}}$. The distribution of $\tilde{\tilde{\mathbf{z}}}_{a_k}$ given σ_{a_k}, $s_{a_{jk}}$, and $\lambda_{a_{jk}}$ is the product of independently normally distributed variables and is represented by

$$p\left(\tilde{\tilde{\mathbf{z}}}_{a_k}\mid\sigma_{a_k},\boldsymbol{\lambda}_{a_k},\mathbf{s}_{a_k}\right)\propto\prod_j\left(\sigma_{a_k}\lambda_{a_{jk}}+s_{a_{jk}}\right)^{-1/2}\exp\left(\frac{-SS_{a_{jk}}/2}{\sigma_{a_k}\lambda_{a_{jk}}+s_{a_{jk}}}\right),\tag{7}$$

where $SS_{a_{jk}} = \left(\tilde{\tilde{z}}_{a_{jk}}\right)^2$.

There are two restrictions on the parameter space of σ_{a_k}. The first restriction states that $\sigma_{a_k}\lambda_{a_{jk}}+s_{a_{jk}}$ is greater than zero for all j, since it represents an eigenvalue of the covariance matrix $\boldsymbol{\Sigma}_{jk}$. A lower bound for σ_{a_k} can be derived, and it follows that $LB_1\left(\sigma_{a_k}\right) = -\min_j s_{a_{jk}}/\lambda_{a_{jk}}$.

The second restriction follows from the positive-definite restriction on the covariance matrix $\boldsymbol{\Sigma}_{jk}$ for all j. The determinant of the covariance matrix can be explicitly derived though the matrix determinant lemma (Appendix B). It follows that

$$LB_2\left(\sigma_{a_k}\right) = -\min_j\left(\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j-\frac{\left(\sum_{i\mid j}\theta_{ij}\right)^2}{1/\sigma_{b_k}+n_j}\right)^{-1}.\tag{8}$$

Then, a shifted-inverse gamma prior can be defined for σ_{a_k}, which integrates the lower bound restrictions. Therefore, let $s_{\sigma_{a_k}} = \min\left(-LB_1\left(\sigma_{a_k}\right),-LB_2\left(\sigma_{a_k}\right)\right)$, and

$$p\left(\sigma_{a_k}\mid s_{\sigma_{a_k}}\right)\propto\left(\sigma_{a_k}+s_{\sigma_{a_k}}\right)^{-g_1-1}\exp\left(-\frac{g_2}{\sigma_{a_k}+s_{\sigma_{a_k}}}\right).\tag{9}$$

The posterior distribution of σ_{a_k} is proportional to the product of the conditional distribution in Equation (7) and the prior in Equation (9), and cannot be obtained in closed form.

4.1.3. Posterior distribution of σ_{b_k}

The restrictions on parameter σ_{b_k} follow in a similar way. The eigenvalue $\sigma_{b_k}\lambda_{b_{jk}}+s_{b_{jk}}$ is greater than zero for all j, which leads to a lower bound for σ_{b_k}, $LB_1\left(\sigma_{b_k}\right) = -\min_j\frac{s_{b_{jk}}}{\lambda_{b_{jk}}}$. The determinant of the covariance matrix needs to be greater than zero, which also restricts the parameter space for σ_{b_k}, and it follows that

$$LB_2\left(\sigma_{b_k}\right) = -\min_j\frac{\left(1+\sigma_{a_k}\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j\right)}{n_j\left(1+\sigma_{a_k}\sum_{i\mid j}\left(\theta_{ij}-\bar{\theta}_j\right)^2\right)}.\tag{10}$$

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

Then, let $s_{\sigma_{b_k}} = \min\left(-LB_1\left(\sigma_{b_k}\right), -LB_2\left(\sigma_{b_k}\right)\right)$, and a prior for σ_{b_k} is a shifted-inverse gamma distribution represented by

$$p\left(\sigma_{b_k} \mid s_{\sigma_{b_k}}\right) \propto \left(\sigma_{b_k} + s_{\sigma_{b_k}}\right)^{-g_1-1} \exp\left(-\frac{g_2}{\sigma_{b_k} + s_{\sigma_{b_k}}}\right). \tag{11}$$

To define the posterior distribution of σ_{b_k}, consider the distribution of $\tilde{\tilde{\mathbf{z}}}_{b_k} = \tilde{\mathbf{w}}_{b_{jk}}^t \tilde{\mathbf{z}}_{jk}$. Let $s_{b_{jk}} = \tilde{\mathbf{w}}_{b_{jk}}^t \tilde{\mathbf{w}}_{b_{jk}}$ and $\lambda_{b_{jk}} = \tilde{\mathbf{w}}_{b_{jk}}^t \boldsymbol{\theta}_j \boldsymbol{\theta}_j^t \tilde{\mathbf{w}}_{b_{jk}}$. The distribution of $\tilde{\tilde{\mathbf{z}}}_{b_k}$ given σ_{b_k}, $s_{b_{jk}}$, and $\lambda_{b_{jk}}$ is represented by

$$p\left(\tilde{\tilde{\mathbf{z}}}_{b_k} \mid \sigma_{b_k}, \boldsymbol{\lambda}_{b_k}, \mathbf{s}_{b_k}\right) \propto \prod_j \left(\sigma_{b_k}\lambda_{b_{jk}} + s_{b_{jk}}\right)^{-1/2} \exp\left(\frac{-SS_{b_{jk}}/2}{\sigma_{b_k}\lambda_{b_{jk}} + s_{b_{jk}}}\right), \tag{12}$$

where $SS_{b_{jk}} = \left(\tilde{\tilde{z}}_{b_{jk}}\right)^2$. It follows that the posterior distribution of σ_{b_k} is proportional to the product of the prior in Equation (11) and the conditional distribution in Equation (12).

5. Bayes factor tests for non-uniform DIF

The BF is used to assess hypotheses concerning the covariance parameters σ_{a_k} and σ_{b_k}. For each covariance parameter, there are three hypotheses of interest. The covariance parameter can be equal to zero, for instance, $H_0 : \sigma_{a_k} = 0$, which represents no DIF in discrimination for item k. The parameter can be greater than zero, $H_2 : \sigma_{a_k} > 0$, which represents DIF in discrimination. A special case can be considered, $H_1 : \sigma_{a_k} < 0$, where DIF is present among group members (instead of DIF between groups). The practical application of this case is not explored in this study. The main purpose of the BF tests is to measure the data support in favor of no (non-uniform) DIF (H_0) and DIF (H_2). An unrestricted hypothesis H_u is used as a reference to examine the preference of each hypothesis in comparison to the unrestricted hypothesis.

The normalizing constant of the marginal likelihood of the (latent) item response data given the covariance parameter in Equations (7) and (12) is unknown. A parameter transformation is used to avoid the computation of a normalizing constant under each hypothesis. It is shown that the marginal likelihood under each hypothesis for the transformed parameter is computationally much more efficient.

5.1. Change of variables

The distribution of $\tilde{\tilde{\mathbf{z}}}_{a_k}$ given σ_{a_k} is a product of independently normally distributed variables with different variance components, but each including σ_{a_k} (Equation (7)). A parameter transformation is considered for which the posterior distribution has a known analytical form. The transformation function $\omega_{a_k} = \frac{\sigma_{a_k}\lambda_{a_{jk}}}{\sigma_{a_k}\lambda_{a_{jk}} + s_{a_{jk}}}$ is defined with inverse function and derivative

$$\sigma_{a_k}\left(\omega_{a_k}\right) = \frac{\omega_{a_k}}{1-\omega_{a_k}}\left(s_{a_{jk}}/\lambda_{a_{jk}}\right)$$

$$\frac{d}{d\omega_{a_k}}\sigma_{a_k}\left(\omega_{a_k}\right) = \frac{s_{a_{jk}}}{\lambda_{a_{jk}}\left(1-\omega_{a_k}\right)^2},$$

respectively. It can be shown that the distribution of the $\tilde{\tilde{\mathbf{z}}}_{a_k}$ given ω_{a_k} is a gamma distribution

$$p\left(\tilde{\tilde{\mathbf{z}}}_{a_k} \mid \omega_{a_k}, \boldsymbol{\lambda}_{a_k}, \mathbf{s}_{a_k}\right) \propto \prod_j \frac{s_{a_{jk}}}{\lambda_{a_{jk}}\left(1-\omega_{a_k}\right)^2}\left(\frac{\omega_{a_k}}{1-\omega_{a_k}}s_{a_{jk}} + s_{a_{jk}}\right)^{-1/2}$$

$$\exp\left(\frac{-SS_{a_{jk}}/2}{\frac{\omega_{a_k}}{1-\omega_{a_k}}s_{a_{jk}} + s_{a_{jk}}}\right)$$

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

$$\propto \prod_j (1-\omega_{a_k})^{1/2} \exp\left(-(1-\omega_{a_k})\, SS_{a_{jk}}/(2s_{a_{jk}})\right)$$
$$= (1-\omega_{a_k})^{J/2} \exp\left(-(1-\omega_{a_k}) \sum_j SS_{a_{jk}}/(2s_{a_{jk}})\right)$$
$$= \tilde{\omega}_{a_k}^{J/2} \exp\left(-\tilde{\omega}_{a_k} \sum_j SS_{a_{jk}}/(2s_{a_{jk}})\right), \tag{13}$$

where $\tilde{\omega}_{a_k} = 1-\omega_{a_k}$, and assume a gamma prior for $\tilde{\omega}_{a_k}$. Then, $\tilde{\omega}_{a_k}$ has a gamma posterior distribution given $\tilde{\mathbf{z}}_{a_k}$ with shape parameter $g_1 + J/2$ and scale parameter $g_2 + \sum_j SS_{a_{jk}}/(2s_{a_{jk}})$. The lower bound for σ_{a_k} is translated to a boundary value for ω_{a_k}, denoted as $UB(\tilde{\omega}_{a_k})$.

In the same way, it can be shown with transformation function $\omega_{b_k} = \frac{\sigma_{b_k}\lambda_{b_{jk}}}{\sigma_{b_k}\lambda_{b_{jk}}+s_{b_{jk}}}$ that $\tilde{\omega}_{b_k} = 1-\omega_{b_k}$ has a gamma distribution. Then, with a gamma prior for the $\tilde{\omega}_{b_k}$, the posterior distribution of $\tilde{\omega}_{b_k}$ is a gamma distribution with shape parameter $g_1 + J/2$ and scale parameter $g_2 + \sum_j SS_{b_{jk}}/(2s_{b_{jk}})$, where $SS_{b_{jk}} = (\tilde{z}_{b_{jk}})^2$. An upper bound for $\tilde{\omega}_{b_k}$, $UB(\tilde{\omega}_{b_k})$, follows from the lower bound restriction on σ_{b_k}

5.2. Bayes factor testing

The hypotheses for σ_{a_k} can be translated to those for $\tilde{\omega}_{a_k}$, taking into account that the upper bound $UB(\tilde{\omega}_{a_k})$ is greater than one ($LB(\sigma_{a_k})$ is less than zero). For hypothesis $H_2(\sigma_{a_k} > 0)$, $\tilde{\omega}_{a_k} \in (-\infty, 1)$, for $H_1(\sigma_{a_k} < 0)$, $\tilde{\omega}_{a_k} \in (1, UB(\tilde{\omega}_{a_k}))$, and for $H_0(\sigma_{a_k} = 0), \tilde{\omega}_{a_k} = 1$. For the unrestricted hypothesis, let $\pi_u(\tilde{\omega}_{a_k})$ be the (unrestricted) gamma prior for $\tilde{\omega}_{a_k}$ with shape g_1 and scale parameter g_2. Following the encompassing prior approach (Böing-Messing & Mulder, 2020; Klugkist & Hoijtink, 2007), for each hypothesis H_t for $t = 0, 1, 2$, the prior under H_t is a restricted version of the encompassing gamma prior, $\pi_t(\tilde{\omega}_{a_k}) = \pi_u(\tilde{\omega}_{a_k})\chi(\tilde{\omega}_{a_k} \mid H_t)$, where

$$\chi(\tilde{\omega}_{a_k} \mid H_t) = \frac{\mathbb{I}(\tilde{\omega}_{a_k} \in H_t)}{P(\tilde{\omega}_{a_k} \mid H_t)}.$$

The BF for the restricted hypothesis H_t for item k concerning $\tilde{\omega}_{a_k}$ against the unrestricted hypothesis H_u can be expressed as

$$BF_{tu}(\tilde{\omega}_{a_k}) = \int_{\mathbb{Z}_k} \left[\int_{-\infty}^{UB(\tilde{\omega}_{a_k})} p(\tilde{\omega}_{a_k} \mid \tilde{\tilde{\mathbf{z}}}_k) \chi(\tilde{\omega}_{a_k} \mid H_t)\, d\tilde{\omega}_{a_k} \right] p(\tilde{\tilde{\mathbf{z}}}_k \mid \mathbf{y}_k)\, d\tilde{\tilde{\mathbf{z}}}_k$$
$$= E\left[\frac{P(\tilde{\omega}_{a_k} \in H_t, \tilde{\omega}_{a_k} < UB(\sigma_{a_k}) \mid \tilde{\tilde{\mathbf{z}}}_k)/P(\tilde{\omega}_{a_k} < UB(\sigma_{a_k}) \mid \tilde{\tilde{\mathbf{z}}}_k)}{P(\tilde{\omega}_{a_k} \in H_t, \tilde{\omega}_{a_k} < UB(\sigma_{a_k}))/P(\tilde{\omega}_{a_k} < UB(\sigma_{a_k}))} \mid \mathbf{y}_k \right], \tag{14}$$

where $\mathbb{Z}_k$ represents the support for $\tilde{\tilde{\mathbf{z}}}_k$. The BF in Equation (14) is expressed as the expected value of a ratio of gamma cumulative distribution functions, which leads to an easy computation of the BF. Furthermore, the BF for two constrained hypotheses also follows easily, since $BF_{tt'} = BF_{tu}/BF_{t'u}$ for $t, t' = 0, 1, 2$. The logarithm of the BF (log-BF) can be computed, where a log-$BF_{tt'}$ of 3–5 is considered strong evidence and above 5 very strong (decisive) evidence in favor of hypothesis t (Kass & Raftery, 1995).

Let $\pi_u(\tilde{\omega}_{b_k})$ denote an unrestricted gamma prior for $\tilde{\omega}_{b_k}$ with shape g_1 and scale parameter g_2. Then, for each hypothesis, the prior under H_t is a restriction on the encompassing gamma prior. Subsequently, the BFs for examining hypotheses about σ_{b_k} follow in a similar way and can also be expressed as a ratio of gamma cumulative distribution functions, while using the appropriate upper bound for $\tilde{\omega}_{b_k}$. The corresponding hypotheses for $\tilde{\omega}_{b_k}$ are given by H_2 $(\sigma_{b_k} > 0)$, $\tilde{\omega}_{b_k} \in (-\infty, 1)$, $H_1(\sigma_{b_k} < 0), \tilde{\omega}_{b_k} \in (1, UB(\tilde{\omega}_{b_k}))$, and $H_0(\sigma_{b_k} = 0), \tilde{\omega}_{b_k} = 1$.

The BF for covariance parameters can be applied to any number of groups and for any number of items. It provides information about which hypothesis is more likely without requiring that any of the hypotheses is true. The BF for testing a precise null hypothesis (H_0) against an unconstrained alternative

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

hypothesis (H_u) is not favoring the null regardless of the information in the data (known as Bartlett's paradox), since the prior probability of $\tilde{\omega}_{b_k} = 1$ is not sensitive to changes in the variance of the gamma prior.

6. Simulation study

For each data replication, cluster sizes were sampled from a uniform distribution between $(n-10, n+10)$ with n the average cluster size and n_j the size of cluster j. The discrimination and difficulty parameters were sampled from a log-normal and normal distribution with mean zero and standard deviation 0.2 and 0.5, respectively. Subject parameters were sampled from a standard normal distribution. Interest was in the covariance parameters, although (latent) cluster means were also estimated. True covariance values for σ_{a_k} and σ_{b_k} were fixed across data replications and are given in the first column of Table 1. The covariances values were set at -0.002 and increased with a step size of 0.02 from item 1 to 10. A positive covariance value represents variance across group-specific item parameters. The grid of covariance values reflects no DIF, low, to a high level of DIF across item parameters. Thus, for instance, item 3, the $\sigma_{a_k} = 0.038$ represents a range of $a_k \pm 0.40$ for the group-specific discriminations. For item 10, it is approximately $a_k \pm 0.84$. The grid of covariance values, specifying the level of DIF, was equal for the discrimination and difficulty parameters.

The BCSM-MI was identified by restricting the scale of the person parameters to have a mean of zero and a variance of one. The prior for each covariance parameter is a shifted-inverse gamma, in Equations (9) and (11), with shape and rate parameter equal to 0.001. The hyper prior parameter settings were equal to those specified in Fox (2024, Appendix C).

Three conditions were considered with 10 items, and $J = 10$, 5, and 2 clusters, and with average cluster sizes $n = 200$, 200, and 250, respectively. A total of 1,000 data replications were made in each condition. Average results are reported in Table 1. The MCMC algorithm was ran for 5,000 iterations with a burn-in period of 1,000 iterations. The effective sample size was at least 400 for each parameter for each data set. The log-BFs were computed of the null hypothesis H_0: $\sigma_{a_k} = 0$ (no DIF) versus the alternative hypothesis H_2: $\sigma_{a_k} > 0$ (DIF), represented as BF_{02}, for all items k. A similar log-BF was computed for σ_{b_k} for all items k.

6.1. Covariance point estimates

From Table 1, for 10 clusters ($J = 10$), it follows that the median estimates (averaged across replications) of the covariance parameters are close to the true values. For small true values, the posterior distribution of the covariance parameters is very skewed to the right. Subsequently, the median point estimate is greater than the mode, which led to a small overestimation of the true value. The average median estimates of σ_{b_k} for assessing uniform DIF were slightly closer to the true values than the median estimates of σ_{a_k} for assessing non-uniform DIF. Higher true values of σ_{a_k} were slightly underestimated.

The (Monte carlo) standard error (SE) increases for increasing true values of the covariance parameters. It is noteworthy that the SEs of the σ_{a_k} are quite comparable to those of the σ_{b_k}. For lower true values, the SE of the σ_{a_k} estimates is slightly higher than those of the σ_{b_k} estimates, where for higher true values, this is reversed. The scale of the difficulties supports more easily large between-cluster differences than the scale of the discriminations. Therefore, it was more difficult to recover the high true values of σ_{a_k} than of σ_{b_k}.

For five clusters—the average total sample size was reduced to 1,000—the average estimates of the covariance parameters show a similar pattern as to those for 10 clusters. However, there is slightly more discrepancy between the true and estimated covariance values. By reducing the number of clusters, there was more sampling variation, and the SEs increased compared to those for 10 clusters. For high true covariance values, the SEs of the σ_{b_k} are higher than those of the σ_{a_k}, since the distribution for generating difficulties had more variance than the one for generating discriminations. The lower bound

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

Table 1. Simulation study: Covariance estimation results for different sample sizes averaged across 1,000 data replications

Item	True	σ_a			σ_b		
	$\sigma_a = \sigma_b$	Median	SE	BF_{02}	Median	SE	BF_{02}
$J = 10,\ n = 200$							
1	−0.002	0.006	0.001	6.939	−0.001	0.001	8.314
2	0.018	0.016	0.003	−0.035	0.014	0.002	1.705
3	0.038	0.027	0.005	−8.122	0.038	0.002	−15.075
4	0.058	0.043	0.004	−21.906	0.055	0.003	−28.832
5	0.078	0.068	0.007	−40.264	0.078	0.003	−49.147
6	0.098	0.080	0.006	−53.921	0.091	0.004	−59.931
7	0.118	0.108	0.007	−74.470	0.117	0.005	−83.296
8	0.138	0.122	0.007	−89.170	0.131	0.005	−97.286
9	0.158	0.126	0.009	−92.674	0.143	0.006	−104.168
10	0.178	0.151	0.008	−114.539	0.184	0.008	−142.483
$J = 5,\ n = 200$							
1	−0.002	0.011	0.004	6.917	0.000	0.002	5.594
2	0.018	0.021	0.007	3.660	0.012	0.004	5.780
3	0.038	0.027	0.007	0.420	0.030	0.005	0.237
4	0.058	0.043	0.009	−4.469	0.048	0.007	−5.795
5	0.078	0.063	0.009	−11.515	0.070	0.008	−13.875
6	0.098	0.058	0.013	−11.462	0.090	0.009	−21.589
7	0.118	0.093	0.012	−23.777	0.114	0.011	−30.506
8	0.138	0.105	0.014	−27.871	0.127	0.012	−35.476
9	0.158	0.093	0.018	−26.394	0.140	0.014	−41.532
10	0.178	0.134	0.018	−40.667	0.234	0.027	−71.162
$J = 2,\ n = 250$							
1	−0.002	0.017	0.053	7.230	0.019	0.059	7.130
2	0.018	0.027	0.092	6.093	0.028	0.084	5.509
3	0.038	0.025	0.078	6.319	0.021	0.062	6.317
4	0.058	0.029	0.090	5.655	0.039	0.096	3.220
5	0.078	0.046	0.104	2.272	0.045	0.099	2.036
6	0.098	0.050	0.114	2.124	0.034	0.092	3.685
7	0.118	0.082	0.125	−3.362	0.079	0.125	−3.308
8	0.138	0.066	0.126	−1.188	0.071	0.124	−2.493
9	0.158	0.082	0.148	−3.561	0.070	0.129	−3.417
10	0.178	0.073	0.137	−1.440	0.237	0.205	−30.487

Note: SE represents the Monte Carlo standard error, and BF_{02} represents the log-BF of no DIF versus (between-cluster) DIF.

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

was closer to zero due to a higher average cluster size. Subsequently, the posterior distributions were more skewed to the right than for 10 clusters.

For two clusters and a total of 500 observations, the SEs increased rapidly and becoming larger than the covariance values. Low and high true covariance values are overestimated for the discrimination and difficulties. For smaller sample sizes, the posterior distribution of the covariance parameter is more skewed to the right, since there is less data information. The posteriors have a longer right tail, which led to a lower performance of the median point estimator.

The mean squared error (MSE) of the discrimination and difficulty estimates show accurate estimation results across all conditions. The MSE slightly increased for increasing covariance values and when decreasing the sample size. The MSEs of the σ_{a_k} are comparable to those of the σ_{b_k}, and only slightly smaller for high true covariance values.

6.2. DIF Bayes factor results

The log-BF was computed for the discriminations and difficulties, where positive (negative) values support the null (alternative) hypothesis in the numerator (denominator). The rows in gray in Table 1 show data evidence in favor of no DIF (null hypothesis) for discriminations and/or difficulties (under the header BF_{02}). When increasing the level of DIF, there is a consistent decrease in the BF_{02} showing more and more support for DIF (alternative hypothesis). This pattern is clearly visible for the discriminations and difficulties. Furthermore, for a decreasing sample size, evidence in favor of DIF was more difficult to achieve. That is, for smaller sample sizes, a higher level of DIF was required to obtain (substantial) evidence in favor of DIF. Decisive evidence of DIF was easier obtained for the difficulty than for the discrimination. For the relatively small sample size of 500 ($J = 2$), the BF_{02} still shows a consistent decrease in value, when the true covariance value increases. Although the point estimates of the true covariances have a relatively high SE, the BF_{02} which is based on the posterior and prior densities shows consistent results.

7. Real data study

Data from the Programme for International Student Assessment or PISA was analyzed for DIF between countries and mode of administration. A total of 14 math items of booklet *B_2015_52* was administered through a computer-based assessment (CBA) in PISA 2015 (OECD, 2016). The same set of items was administered within "standard" paper-based assessment (PBA) in 2012 in booklet *B_2012_4*. The item responses from Belgium (BEL), Germany (DEU), and The Netherlands (NED) were considered. The response patterns of 249 (BEL), 148 (DEU), and 183 (NED) respondents were collected from a PBA in 2012, and of 677 (BEL), 338 (DEU), and 351 (NED) respondents from a CBA in 2015.

An overview of assessment and score differences between the 2012 and 2015 cycle is given by Jerrim (2016). It was concluded that pupils who did the computer-based test under performed to those who did the paper-based test. The mode of test administration had an impact on the pupils' scores. For several items, there was a mode effect in which, for instance, items appeared to be less difficult in a PBA than in a CBA. The BCSM-MI method was used to examine country-specific item–mode interactions and country–item-specific person–mode interactions, while accounting for latent group means and variances. The BCSM-MI did not require any anchor items, which enabled a simultaneous DIF analysis of the 14 items across the three selected countries, each with two modes of administration.

In Table 2, for each item parameter, the covariance value, SE, and the level of BF_{02} (no DIF versus between-cluster DIF) are given. For item discrimination 10, the BF_{02} shows decisive evidence in favor of DIF. Although the BF_{02} is also negative for item discrimination 7, it shows only weak evidence in favor of DIF. For item difficulties 1, 2, and 12, there is also decisive evidence in favor of DIF. For the item parameters associated with DIF, the size of the BF_{02} is associated with the level of covariance, where more negative BF_{02} values correspond to higher covariance values. According to the BF results,

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

Table 2. PISA 2012–2015: DIF examination of paper-based and computed-based assessed math items across three countries

Item		BCSM-MI			MG-IRT				
	Median	SE	BF_{02}	*LR*	$P(LR)$	LR_{seq}	$P(LR_{seq})$	MD	RMSD
Discrimination σ_a									
1	0.011	0.002	8.104	5.576	0.350				
2	0.008	0.001	8.871	2.912	0.713				
3	0.032	0.005	3.033						
4	0.009	0.001	8.496						
5	0.016	0.001	6.907						
6	0.015	0.001	7.691						
7	0.030	0.003	−1.250	8.815	0.117				
8	0.023	0.003	7.537						
9	0.038	0.004	3.335						
10	0.115	0.007	−33.694	34.445	0.000	34.221	0.000		
11	0.012	0.001	7.898						
12	0.025	0.003	3.457	4.110	0.027	11.998	0.035		
13	0.011	0.001	8.527						
14	0.013	0.003	8.389						
Difficulty σ_b									
1	0.111	0.007	−34.109	67.065	0.000	67.065	0.000	0.141	0.120
2	0.064	0.009	−10.508	19.301	0.000	19.301	0.002	0.069	0.020
3	0.021	0.002	5.462						
4	0.004	0.000	9.259						
5	0.010	0.001	7.494						
6	0.016	0.002	8.505						
7	0.022	0.001	1.908	17.928	0.003	17.928	0.003	0.089	0.078
8	0.015	0.001	6.240						
9	0.018	0.002	6.818						
10	0.011	0.001	8.386	3.159	0.675			0.019	0.035
11	0.006	0.001	8.902						
12	0.060	0.003	−11.973	31.068	0.000	31.068	0.000	0.086	0.045
13	0.007	0.001	8.897						
14	0.011	0.002	7.801						

Note: SE represents the Monte Carlo standard error. *LR*: Likelihood-ratio test. $P(LR)$: *P*-value LR test. LR_{seq} (sequential) final LR test. $P(LR_{seq})$: *P*-value (sequential) final LR test, MD: mean deviation, and RMSD: root mean square deviation.

those items (rows) shown in light gray have a negative BF_{12} for at least one item parameter, and item parameters in dark gray show DIF.

The items for which the BF_{12} was positive for discrimination and difficulty were considered invariant and used as anchors (i.e., free of DIF) in an MG-IRT analysis. A baseline MG-IRT model was defined

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

using the anchors for an LR test procedure (Millsap, 2011). The latent mean and variance parameters were considered free parameters, given a reference group for which the latent mean and variance were restricted to 0 and 1, respectively. Items 1, 2, 7, 10, and 12 were examined for DIF in discrimination and difficulty, which are colored light gray in Table 2. The R-package *mirt* (Chalmers, 2012) was used to perform LR testing for DIF by comparing the baseline model to various constrained MG-IRT models, in which item parameters were constrained to be equal across groups.

For each item examined for DIF, an MG-IRT model with the discrimination parameter constrained was compared to the baseline model using the LR test. In a similar way, MG-IRT models were defined for which each of the item difficulty parameters was restricted. Thus, a total of 10 models were needed to examine each item parameter of the five examined items. In Table 2, the LR values and corresponding *P*-values for the discriminations and difficulties are given under the header LR and P(LR), respectively. High LR values correspond to negative BF_{02} values and small *P*-values. There appears to be an agreement between the methods about which item parameters show DIF. However, the difficulty parameter of item 7 appears to show DIF according to the LR test, while the BF test shows weak support for no DIF. For the discrimination parameter of item 7, the BF_{02} shows weak evidence for DIF, while the LR test does not. However, note that the log-BF test is a simultaneous DIF analysis of all items and is not based on a comparison(s) with a baseline model.

Subsequently, DIF in discriminations and difficulties was each sequentially tested under the MG-IRT model. In this sequential procedure, items were restricted to be invariant when satisfying the AIC criterion, and the remaining set of items was examined until no more invariant items were found. In Table 2, for the final set of item parameters that displayed DIF the LR statistic and *P*-value are reported, which are denoted as LR_{seq} and $P(LR_{seq})$, respectively. For the discriminations, items 10 and 12 are marked as DIF, and for the difficulties, items 1, 2, 7, and 12. However, when using the sample-adjusted BIC as a selection criterion to decide on which item is restricted to be invariant, the LR sequential test results changed. In that case, only the discrimination parameter of item 10 is still marked to show DIF, while the discrimination of item 12 is considered to be free of DIF. This is in line with the BCSM-MI analysis. Furthermore, only the difficulties of items 1 and 12 show DIF, and difficulties of items 2 and 7 do not show DIF. This is not in line with the BCSM-MI results, where the difficulty of item 2 also showed DIF.

The mean deviation (MD) and the root mean square deviation (RMSD) were computed for each item (Köhler et al., 2024, Equations (2) and (3)). In a baseline MG-IRT model, all item parameters were restricted across groups to estimate the common set of item parameters. Subsequently, group-specific item parameters were estimated using the anchors from the BCSM-MI analysis. In Table 2, for the examined set of items, the maximum MD and RMSD across groups are represented in the rows of the difficulty parameters. The MD is considered to assess the difference in item difficulty between the group specific item response function and the common item response function, where the RMSD measures this (squared) difference with respect to difficulty and discrimination. The estimated levels of MD are in correspondence with the BF. Items 1, 2, and 12 have relatively high MD values corresponding with high BF values for the difficulty parameters. Item 7 has a relatively high MD value for which the RMSD value is also relatively high, which corresponds to the weak support of the BF for DIF in discrimination.

In the PISA study of Köhler et al. (2024), under a two-parameter IRT model, an item's response function is considered to differ across groups, when the RMSD and the absolute value of MD are greater than 0.12. According to that criterion, only item 1 is considered to show DIF. It could be that for large sample sizes comparable to PISA studies, this threshold leads to acceptable Type-1 errors. Obviously, a high threshold will reduce the number of false positives. However, for item 1, an estimated difficulty covariance of 0.111 indicates that the range of difficulty values across groups is around ±0.67. An RMSD value greater than 0.15 to identify items behaving differently across groups was used in von Davier et al. (2019). When using this threshold, none of the items would be classified to show DIF. Note that the simulation study showed that for five groups with a group size of around 200, a discrimination covariance of 0.06, and a difficulty covariance of 0.06, already led to strong evidence in favor of DIF. Therefore, the item response function of item 10 differs across groups with respect to the discrimination.

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

The item response functions of items 1, 2, and 12 also differ across groups with respect to the item difficulty. For these item response functions, the log-BF shows very strong evidence in favor of DIF, and the very high LR values also lead to rejecting the null hypothesis of a constrained (invariant) item parameter.

8. Discussion

The BCSM-MI is proposed for non-uniform DIF testing for any number of groups and unbalanced test designs. All item parameters can be examined simultaneously for DIF without needing any anchor items or a sequential test procedure. The method does not require selecting any anchor items, before DIF testing a set of items. This avoids the problem that DIF results from a sequential procedure depends on the set of anchor items—the specification of a baseline model—and on the order in which the items are sequentially tested.

The BCSM-MI does not include group-specific item parameters, which relaxes restriction on the sample size to obtain stable model parameter estimates. The DIF for each item parameter is represented by a single covariance parameter, and an increasing level of covariance is associated with an increase in the level of DIF across groups. In an MG-IRT model, partial item invariance across groups can be modeled by including group-specific item parameters. In the PISA study with many countries, the number of group-specific item parameters can increase rapidly, which can also lead to scaling problems due to a lack of invariant items. Therefore, groups showing a similar level of item misfit receive the same (unique) group-specific item parameter, to restrict the total set of item parameters to be calibrated (von Davier et al., 2019). The BCSM-MI uses a single covariance parameter for each item parameter to allow for DIF independent on the number of groups. From a fitted BCSM-MI, group-specific item parameters can be computed for items showing DIF. This would lead to a calibrated set of common and unique item parameters without specifying anchor items beforehand.

In Section S2 of the Supplementary Material, for the simulation design discussed in Section 6, superiority of the BCSM-MI over regularized DIF is shown with respect to the true and false positive rates.

In the PISA study, the BCSM-MI models DIF in discrimination and difficulty. Therefore, latent variable estimates are free of (non)-uniform DIF with respect to clusters defined by countries and mode of administration. The BCSM-MI can be used to detect DIF, but also to compute latent variable estimates corrected for possible DIF bias. Subsequently, country-specific latent variable estimates and between-country and between-mode of administration differences are corrected for DIF effects. Post-hoc (cluster-specific) DIF effects can be computed under the BCSM-MI, which can provide further information about which clusters differ with respect to the functioning of items.

The shifted-inverse gamma prior is defined for each covariance parameter to allow for negative as well as positive covariance values, while accounting for the positive-definite restriction of the covariance matrix, which implies a lower bound restriction on the covariance parameter. More research is needed to examine the impact of the shifted-inverse gamma prior on the shape of the posterior. The posterior distribution of the covariance parameter is often skewed to the right, which makes it more difficult to estimate the mode. The simulation results showed that 95% highest posterior density (HPD) intervals had the tendency to include small covariance values (associated with high posterior density) and exclude high covariance values (associated with low posterior density). This led to conservative coverage rates, especially for true small covariance values. Subsequently, evaluation of the DIF hypotheses through posterior probability computing usually favored the DIF hypotheses and disfavored the non-DIF hypotheses, since most of the posterior mass was located above zero covariance. This motivated the use of the BF for DIF testing, because it is based on (marginal) likelihoods to make inferences about the DIF hypotheses. The (conjugate) gamma prior for the ω parameter is an encompassing prior, which facilitates much easier computation of the BF. More research is needed to examine the influence of the prior on the BF results, and performance differences of the BF under different priors.

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10101.

Data availability statement. The R-package *BCSMMI* for Bayesian covariance modeling of DIF can be found on www.Jean-PaulFox.com under Software.

Funding statement. This research received no grant funding from any funding agency, commercial or not-for-profit sectors. Open access funding provided by University of Twente.

Competing interests. The author declares none.

References

Belzak, W., & 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. https://doi.org/10.1037/met0000253

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. van der, Linden, & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). Springer.

Böing-Messing, F., & Mulder, J. (2020). Bayes factors for testing order constraints on variances of dependent outcomes. *The American Statistician*, *75*(2), 152–161. https://doi.org/10.1080/00031305.2020.1715257

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, Y., Li, C., Ouyang, J., & Xu, G. (2023). DIF statistical inference without knowing anchoring items. *Psychometrika*, *88*(4), 1097–1122.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., & Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, *5*, 329–359. https://doi.org/10.1007/BF02124750

De Jong, M., Steenkamp, J. E. M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research*, *34*, 260–278. https://doi.org/10.1086/518532

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*, 278–295. https://doi.org/10.1177/0146621605275728

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Fox, J.-P. (2024). Redefining item response models for small samples. *Journal of Educational and Behavioral Statistics*, *50*(2), 272–295. https://doi.org/10.3102/10769986241269886

Fox, J.-P., Koops, J., Feskens, R., & Beinhauer, L. (2020). Bayesian covariance structure modelling for measurement invariance testing. *Behaviormetrika*, *47*(2), 385–410. https://doi.org/10.1007/s41237-020-00119-3

Fox, J.-P., Mulder, J., & Sinharay, S. (2017). Bayes factor covariance testing in item response models. *Psychometrika*, *82*(4), 979–1006. https://doi.org/10.1007/s11336-017-9577-6

Fox, J.-P., & Smink, W. A. C. (2023). Assessing an alternative for "negative variance components": A gentle introduction to Bayesian covariance structure modeling for negative associations among patients with personalized treatments. *Psychological Methods*, *28*(1), 1–20. https://doi.org/10.1037/met0000442

Hartig, J., Köhler, C., & Naumann, A. (2020). Using a multilevel random item Rasch model to examine item difficulty variance between random groups. *Psychological Test and Assessment Modeling*, *62*, 11–27.

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, *23*(4), 495–518. https://doi.org/10.1080/0969594x.2016.1147420

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*, 6367–6379. https://doi.org/10.1016/j.csda.2007.01.024

Köhler, C., Hartig, J., Khorramdel, L., & Pokropek, A. (2024). DIF detection for multiple groups: Comparing three-level GLMMs and multiple-group IRT models. *Journal of Educational Measurement*, *61*, 303–324. https://doi.org/10.1111/jedm.12384

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis Group.

Muthèn, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, *10*, 121–132. https://doi.org/10.2307/1164839

Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*, 257–274. https://doi.org/10.1177/014662169602000306

Nielsen, N. M., Smink, W. A. C., & Fox, J.-P. (2021). Small and negative correlations among clustered observations: Limitations of the linear mixed effects model. *Behaviormetrika*, *48*, 51–77. https://doi.org/10.1007/s41237-020-00130-8

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

OECD. (2016). *PISA 2015 Techical report*. OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/
Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods*, *52*(1), 279–294.
Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*, 184–199. https://doi.org/10.1177/0146621619835496
Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer, & H. Braun (Eds.), *Test validity* (pp. 147–170). Erlbaum.
Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*, 443–464. https://doi.org/10.3102/10769986030004443
Verhagen, J., & Fox, J.-P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383–401. https://doi.org/10.1111/j.2044-8317.2012.02059.x
von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 466–488. https://doi.org/10.1080/0969594x.2019.1586642
Zhang, D., Khalili, A., & Asgharian, M. (2022). Post-model-selection inference in linear regression models: An integrated review. *Statistic Surveys*, *16*, 86–136. https://doi.org/10.1214/22-SS135

Appendix A: Conditional expectation latent response

The expression for the conditional expectation of the latent response in Equation (5) is derived from the multivariate normal distribution. The multivariate normal distribution of the latent responses in group j has structured covariance matrix $\mathbf{\Sigma}_{jk}$ according to Equation (3). Consider $\sigma_{b_k} = 0$, then the inverse of the covariance matrix is given by

$$\begin{aligned}\mathbf{\Sigma}_{jk}^{-1} &= \left(\mathbf{I}_{n_j} + \sigma_{a_k}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t\right)^{-1} \\ &= \mathbf{I}_{n_j} - \frac{\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t}{\lambda_j},\end{aligned}$$

where $\lambda_j = 1/\sigma_{a_k} + \sum_s \theta_{sj}^2$. Using this expression for the covariance matrix and $\mu_{ijk} = a_k\theta_{ij} - b_k$, the multivariate normal density of $\mathbf{Z}_{jk}$ can be expressed as

$$\begin{aligned}\log p\left(\mathbf{Z}_{jk}\right) &\propto \frac{-1}{2}\left(\mathbf{Z}_{jk} - \boldsymbol{\mu}_{jk}\right)^t\left(\mathbf{I}_{n_j} - \frac{\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t}{\lambda_j}\right)\left(\mathbf{Z}_{jk} - \boldsymbol{\mu}_{jk}\right) \\ &\propto \frac{-1}{2}\sum_i\left(Z_{ijk} - \mu_{ijk}\right)^2 + \left(\sum_i\sum_{s\neq i}\frac{\theta_{ij}\theta_{sj}}{\lambda_j}\left(Z_{ijk} - \mu_{ijk}\right)\left(Z_{sjk} - \mu_{sjk}\right)\right. \\ &\quad \left. + \frac{1}{2}\sum_i\frac{\theta_{ij}^2}{\lambda_j}\left(Z_{ijk} - \mu_{ijk}\right)^2\right) \\ &\propto \frac{-1}{2}\sum_i\left(\left(Z_{ijk} - \mu_{ijk}\right) - \frac{\theta_{ij}}{\lambda_{ij}}\sum_{s\neq i}\theta_{sj}\left(Z_{sjk} - \mu_{sjk}\right)\right)^2 / \left(\frac{\lambda_j}{\lambda_{ij}}\right),\end{aligned}$$

where $\lambda_{ij} = 1/\sigma_{a_k} + \sum_{s\neq i}\theta_{sj}^2$. The expression of the multivariate normal density reveals that it consists of a product of conditionally normally distributed variables Z_{ijk} given $\mathbf{Z}_{(-i)jk}$, with conditional expected values

$$E\left(Z_{ijk} \mid \mathbf{Z}_{(-i)jk}\right) = a_k\theta_{ij} - b_k + \theta_{ij}\sum_{s\neq i}\frac{\theta_{sj}\left(Z_{sjk} - \mu_{sjk}\right)}{1/\sigma_{a_k} + \sum_{s\neq i}\theta_{sj}^2},$$

which represents the expression in Equation (5).

Appendix B: Restrictions on the parameter space

An analytical expression for the inverse of the covariance matrix is obtained using the Sherman–Morrison formula. This formula is applied by partitioning the covariance matrix in the sum of an invertible matrix $\mathbf{A}_{jk}$ and the outer product of $\boldsymbol{\theta}_j$.

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

Then, the inverse of the matrix $\mathbf{A}_{jk}$ is derived by applying once more the Sherman–Morrison formula. This way an analytical expression for the inverse is obtained

$$\begin{aligned}\boldsymbol{\Sigma}_{jk}^{-1} &= \left(\left(\mathbf{I}_{n_j} + \sigma_{b_k}\mathbf{J}_{n_j}\right) + \sigma_{a_k}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t\right)^{-1} \\ &= \left(\mathbf{A}_{jk} + \sigma_{a_k}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t\right)^{-1} \\ &= \left(\mathbf{A}_{jk}^{-1} - \frac{\mathbf{A}_{jk}^{-1}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t\mathbf{A}_{jk}^{-1}}{1/\sigma_{a_k} + \boldsymbol{\theta}_j^t\mathbf{A}_{jk}^{-1}\boldsymbol{\theta}_j}\right),\end{aligned} \tag{B1}$$

and subsequently, the $\mathbf{A}_{jk}^{-1}$ is given by

$$\begin{aligned}\mathbf{A}_{jk}^{-1} &= \left(\mathbf{I}_{n_j} + \sigma_{b_k}\mathbf{J}_{n_j}\right)^{-1} \\ &= \left(\mathbf{I}_{n_j} - \frac{\mathbf{J}_{n_j}}{1/\sigma_{b_k} + n_j}\right).\end{aligned}$$

Note that the inverse of the covariance matrix for any partitioned response set, in which, for instance, a person i is excluded, can be derived in a similar way.

The determinant of the covariance matrix can be explicitly derived though the matrix determinant lemma. It follows that

$$\begin{aligned}\det\left(\boldsymbol{\Sigma}_{jk}\right) &= \det\left(\mathbf{I}_{n_j} + \sigma_{b_k}\mathbf{J}_{n_j} + \sigma_{a_k}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t\right) \\ &= \det(\mathbf{I}_{n_j} + \sigma_{b_k}\mathbf{J}_{n_j})\left(1 + \sigma_{a_k}\boldsymbol{\theta}_j^t\left(\mathbf{I}_{n_j} + \sigma_{b_k}\mathbf{J}_{n_j}\right)^{-1}\boldsymbol{\theta}_j\right) \\ &= \det(\mathbf{I}_{n_j})\left(1 + \sigma_{b_k}n_j\right)\left(1 + \sigma_{a_k}\boldsymbol{\theta}_j^t\left(\mathbf{I}_{n_j} + \sigma_{b_k}\mathbf{J}_{n_j}\right)^{-1}\boldsymbol{\theta}_j\right) \\ &= \left(1 + \sigma_{b_k}n_j\right)\left(1 + \sigma_{a_k}\left(\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j - \frac{\left(\sum_{i|j}\theta_{ij}\right)^2}{1/\sigma_{b_k} + n_j}\right)\right) \\ &= \left(1 + \sigma_{b_k}n_j\right)\left(1 + \sigma_{a_k}\left(\frac{\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j/\sigma_{b_k} + n_j\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j - n_j^2\bar{\theta}_j^2}{1/\sigma_{b_k} + n_j}\right)\right) \\ &= \left(1 + \sigma_{b_k}n_j\right)\left(1 + \sigma_{a_k}\left(\frac{\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j + \sigma_{b_k}n_j\sum_{i|j}\left(\theta_{ij} - \bar{\theta}_j\right)^2}{1 + \sigma_{b_k}n_j}\right)\right) \\ &= \left(1 + \sigma_{b_k}n_j\right) + \sigma_{a_k}\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j + \sigma_{a_k}\sigma_{b_k}n_j\sum_{i|j}\left(\theta_{ij} - \bar{\theta}_j\right)^2.\end{aligned}$$

The determinant of the covariance matrix $\boldsymbol{\Sigma}_{jk}$ needs to be greater than zero, which leads to restrictions on the parameter space of σ_{a_k} and σ_{b_k}. For σ_{b_k}, the following lower bound is derived:

$$LB_2\left(\sigma_{b_k}\right) = \frac{-\left(1 + \sigma_{a_k}\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j\right)}{n_j\left(1 + \sigma_{a_k}\sum_{i|j}\left(\theta_{ij} - \bar{\theta}_j\right)^2\right)}.$$

For σ_{a_k}, the following lower bound is derived:

$$LB_2\left(\sigma_{a_k}\right) = -\min_j\left(\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j - \frac{\left(\sum_{i|j}\theta_{ij}\right)^2}{1/\sigma_{b_k} + n_j}\right)^{-1}.$$

In the sampling step for α_{jk}, the expression of the posterior variance $\Psi_{a_{jk}}$ in Equation (C3) is restricted to be positive. This also implies the restriction that $\sigma_{b_k} > -1/n_j$. This does not lead to a change in the lower bound for σ_{a_k}. The variance $\Psi_{a_{jk}}$ is restricted to be positive, which leads to the same lower bound restriction for σ_{a_k}:

$$\begin{aligned}LB_2\left(\sigma_{a_k}\right) &= -\min_j\left(\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j - \frac{\left(\boldsymbol{\theta}_j^t\mathbf{1}_{n_j}\right)^2}{1/\sigma_{b_k} + n_j}\right)^{-1} \\ &= -\min_j\left(\frac{\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j + \sigma_{b_k}n_j\sum_{i|j}\left(\theta_{ij} - \bar{\theta}\right)^2}{1 + \sigma_{b_k}n_j}\right)^{-1}.\end{aligned}$$

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

When $\sigma_{b_k} > \frac{-1}{n}$, the denominator is restricted to be positive. Furthermore, the numerator is also restricted to be positive, since $\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j > \sum_{i|j}\left(\theta_{ij} - \overline{\theta}_j\right)^2$.

Appendix C: MCMC algorithm

The steps of the MCMC scheme are described. For the (shifted) inverse gamma prior, general prior parameters g_1 and g_2 are used, which can be adjusted in each step.

- The latent item response Z_{ijk} is conditionally sampled given the remaining latent responses in cluster j and the observed item response Y_{ijk}. The conditional distribution of the latent response is truncated normal with mean

$$\mu_{z_{ijk}} = \left(a_k\theta_{ij} - b_k\right) + \boldsymbol{\Sigma}_{(i,-i)jk}\boldsymbol{\Sigma}^{-1}_{(-i,-i)jk}\left(\mathbf{z}_{(-i)jk} - \left(a_k\boldsymbol{\theta}_{(-i)j} - b_k\right)\right) \tag{C1}$$

and variance

$$\sigma_{z_{ijk}} = \boldsymbol{\Sigma}_{(i,i)jk} - \boldsymbol{\Sigma}_{(i,-i)jk}\boldsymbol{\Sigma}^{-1}_{(-i,-i)jk}\boldsymbol{\Sigma}_{(-i,i)jk}, \tag{C2}$$

where the indices $(-i)$ represent all i in group j excluding i, and $Z_{ijk} \geq 0$ when $Y_{ijk} = 1$, and $Z_{ijk} < 0$ otherwise. The inverse of the covariance matrix is obtained by applying the Sherman–Morrison formula (Appendix B). An analytical expression for the inverse of the covariance matrix excluding row and column i, $\boldsymbol{\Sigma}^{-1}_{(-i,-i)jk}$, can be obtained in a similar way.
- Let $\theta_{ij} \sim \mathcal{N}\left(\mu_{\theta_j}, \sigma_{\theta_j}\right)$. When $\sigma_{a_k} > 0$, group-specific discrimination parameters a_{jk} are sampled. When $\sigma_{a_k} \leq 0$, the $a_{jk} = a_k$ for all j. For $\sigma_{a_k} > 0$, the posterior distribution of the group-specific discrimination parameter a_{jk} is normal with variance

$$\Psi_{a_{jk}} = \left(\left(\boldsymbol{\theta}_j^t\boldsymbol{\theta}_j - \frac{\left(\boldsymbol{\theta}_j^t\mathbf{1}_{n_j}\right)^2}{1/\sigma_{b_k} + n_j}\right) + 1/\sigma_{a_k}\right)^{-1} \tag{C3}$$

and mean

$$\mu_{a_{jk}} = \Psi_{a_{jk}}\left(\left(\boldsymbol{\theta}_j^t\left(\mathbf{z}_{jk} + b_k\right) - \frac{\left(\boldsymbol{\theta}_j^t\mathbf{1}_{n_j}\right)\left(\mathbf{1}_{n_j}^t\left(\mathbf{z}_{jk} + b_k\right)\right)}{1/\sigma_{b_k} + n_j}\right) + \frac{a_k}{\sigma_{a_k}}\right). \tag{C4}$$

The expression for the variance in Equation (C3) restricts the σ_{b_k} to be greater than $-1/n_j$. Otherwise, the variance term $\Psi_{a_{jk}}$ is no longer restricted to be positive (see Appendix B). This lower bound is slightly above the lower bound in Equation (10) to ensure a positive-definite covariance matrix. Therefore, this lower bound is only implied in the sampling step for the θ_{ij}. The posterior distribution of $\boldsymbol{\theta}$ given the group-specific item discriminations, $\mathbf{a}_j$, is multivariate normal with covariance matrix

$$\boldsymbol{\Psi}^{-1}_{\theta_j} = \left(\sum_k \alpha^2_{jk}\left(\mathbf{I}_{n_j} - \frac{\mathbf{J}_{n_j}}{1/\sigma_{b_k} + n_j}\right)\right) + \sigma^{-1}_{\theta_j}\mathbf{I}_{n_j}$$

and mean

$$\boldsymbol{\mu}_\theta = \boldsymbol{\Psi}^{-1}_{\theta_j}\left(\sum_k \alpha_{jk}\left(\left(\mathbf{z}_{jk} + b_k\right) - \frac{\sum_{i|j} z_{ijk} + b_k}{1/\sigma_{b_k} + n_j}\right) + \mu_{\theta_j}/\sigma_{\theta_j}\right). \tag{C5}$$

This procedure is used in the simulation study (Section 6) and shows good performance. However, it ignores the impact of a negative σ_{a_k}. In Appendix D, it is shown how a Metropolis–Hastings step can be defined to account for negative covariances. The posterior sampling steps for μ_{θ_j} and σ_{θ_j} can be found in Fox (2010, Chapter 7).
- The (invariant) discrimination parameters a_k are sampled from a normal distribution with variance

$$\Psi_{a_k} = \left(\sum_j \boldsymbol{\theta}_j^t\boldsymbol{\Sigma}^{-1}_{jk}\boldsymbol{\theta}_j + 1/\sigma^2_a\right)^{-1} \tag{C6}$$

and mean

$$\mu_{a_k} = \Psi_{a_k}\left(\sum_j \boldsymbol{\theta}_j^t\boldsymbol{\Sigma}^{-1}_{jk}\left(\mathbf{z}_{jk} + b_k\right) + \mu_a/\sigma^2_a\right). \tag{C7}$$

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

- The (invariant) difficulty parameters b_k are sampled from a normal distribution with variance

$$\Psi_{b_k} = \left(\sum_j \mathbf{1}_j^t \boldsymbol{\Sigma}_{jk}^{-1} \mathbf{1}_j + 1/\sigma_b^2 \right)^{-1} \tag{C8}$$

and mean

$$\mu_{b_k} = \Psi_{b_k} \left(\sum_j \mathbf{1}_j^t \boldsymbol{\Sigma}_{jk}^{-1} \left(\mathbf{z}_{jk} - a_k \boldsymbol{\theta}_j \right) + \mu_b / \sigma_b^2 \right). \tag{C9}$$

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

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

respectively.
- The prior parameters μ_b and variance σ_b^2 have a normal inverse-gamma hyper prior with $\sigma_b^2 \sim \mathcal{IG}(g_1, g_2)$ and $\mu_b \sim N\left(\mu_{b_0}, \sigma_b^2 / n_{b_0}\right)$. Then, the posterior distribution of μ_b and variance σ_b^2 is given by

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

respectively.
- The sampling of covariance parameters σ_{a_k} is done with slice sampling. Therefore, a uniformly distributed value u under the posterior density of $\sigma_{a_k}^{(m-1)}$ is sampled, where $\sigma_{a_k}^{(m-1)}$ is retrieved in iteration $m-1$. Then, $\sigma_{a_k}^{(m)}$ can be uniformly sampled conditional on the sampled u. That is,

$$u_j \sim \mathcal{U}\left(0, p\left(\sigma_{a_k}^{(m-1)} \mid \tilde{\tilde{z}}_{a_{jk}}, \lambda_{a_{jk}}, s_{a_{jk}}\right)\right) \tag{C10}$$
$$\sigma_{a_k}^{(m)} \sim \mathcal{U}\left(S_a(u)\right), \tag{C11}$$

where $S_a(u) = \bigcap_j S_{a_j}(u_j)$ and $S_{a_j}(u_j) = \left\{\sigma_{a_k} : p\left(\sigma_{a_k} \mid \tilde{\tilde{z}}_{a_{jk}}, \lambda_{a_{jk}}, s_{a_{jk}}\right) > u_j\right\}$. The set $S_a(u)$ can be obtained using the Lambert W function (Corless et al., 1996). An upper bound can be placed on the parameter space of σ_{a_k}, and this upper bound is easily integrated in the target space $S_a(u)$ from which $S_a(u)$ is sampled.
- The covariance parameter σ_{b_k} is sampled using slice sampling. A uniformly distributed value u under the posterior density of $\sigma_{b_k}^{(m-1)}$ is sampled, where $\sigma_{b_k}^{(m-1)}$ is retrieved in iteration $m-1$. The $\sigma_{a_k}^{(m)}$ is uniformly sampled given u

$$u_j \sim \mathcal{U}\left(0, p\left(\sigma_{b_k}^{(m-1)} \mid \tilde{\tilde{z}}_{b_{jk}}, \lambda_{b_{jk}}, s_{b_{jk}}\right)\right) \tag{C12}$$
$$\sigma_{b_k}^{(m)} \sim \mathcal{U}\left(S_b(u)\right), \tag{C13}$$

where $S_b(u) = \bigcap_j S_{b_j}(u_j)$ and $S_{b_j}(u_j) = \left\{\sigma_{b_k} : p\left(\sigma_{b_k} \mid \tilde{\tilde{z}}_{b_{jk}}, \lambda_{b_{jk}}, s_{b_{jk}}\right) > u_j\right\}$. An upper bound can also be placed on the parameter space of σ_{b_k}, which can be integrated in the allowable set $S_b(u)$.

Appendix D: Sampling person parameters

The sampling of the person parameters is complicated, since the person parameters are also involved in the structured covariance matrix of the BCSM-MI. When $\sigma_{a_k} > 0$, and assuming a normal distribution for the random discrimination parameters with mean a_k, the BCSM-MI can be redefined with a random discrimination parameter in the mean term. It follows that

$$\mathbf{Z}_{jk} = \mathcal{N}\left(a_{jk}\boldsymbol{\theta}_j - b_k, \boldsymbol{\Sigma}_{j_b}\right), \tag{D1}$$

where $a_{jk} \sim \mathcal{N}(a_k, \sigma_{a_k})$, $\boldsymbol{\theta}_j \sim \mathcal{N}\left(0, \sigma_{\theta_j} \mathbf{I}_{n_j}\right)$, and $\boldsymbol{\Sigma}_{j_b} = \mathbf{I}_{n_j} + \sigma_{b_k} \mathbf{J}_{n_j}$.

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press

When at least one $\sigma_{a_k} < 0$, a Metropolis–Hastings step can be defined to draw a value for θ_{ij}. Let the set $\Omega_j = \{s \mid \sigma_{a_s} > 0\}$ represent the items s for which $\sigma_{a_s} \geq 0$. Subsequently, let $\mathbf{z}_{js}(s \in \Omega_j)$ represent the corresponding item responses in group j. Then, group-specific discriminations a_{js} are sampled for all items $s \in \Omega_j$. A candidate value for $\boldsymbol{\theta}_j^*$ is sampled given the group-specific discriminations a_{js}, according to the model specification in Equation (D1). The candidate value is compared to the former value using the following acceptance ratio:

$$R\left(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j\right) = \frac{p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j^*, \boldsymbol{\Sigma}_j\right) \prod_{s \in \Omega_j} p\left(\mathbf{z}_{js} \mid \boldsymbol{\theta}_j, a_{js}, \boldsymbol{\Sigma}_{j_b}\right) p\left(a_{js} \mid \boldsymbol{\theta}_j^*, \mathbf{z}_{js}, \boldsymbol{\Sigma}_{j_b}\right)}{p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_j\right) \prod_{s \in \Omega_j} p\left(\mathbf{z}_{js} \mid \boldsymbol{\theta}_j^*, a_{js}, \boldsymbol{\Sigma}_{j_b}\right) p\left(a_{js} \mid \boldsymbol{\theta}_j, \mathbf{z}_{js}, \boldsymbol{\Sigma}_{j_b}\right)}. \tag{D2}$$

When $\sigma_{a_k} < 0$ for all k, a candidate value can be generated by assuming that $a_{kj} = a_k$ for all j and k.

The acceptance ratio in Equation (D2) equals one, when $\sigma_{a_k} > 0$ for all k. In that case, the terms conditioning on $\boldsymbol{\theta}_j$ and those on $\boldsymbol{\theta}_j^*$ cancel out. The distribution of the (latent) item response data for group j is expressed as follows:

$$p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_j\right) = \frac{p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j, \mathbf{a}_j, \boldsymbol{\Sigma}_{j_b}\right) p\left(\mathbf{a}_j \mid \sigma_a\right)}{p\left(\mathbf{a}_j \mid \boldsymbol{\theta}_j, \mathbf{z}_j, \boldsymbol{\Sigma}_{j_b}\right)}, \tag{D3}$$

which is allowed when $\sigma_{a_k} > 0$ for all k. The acceptance ratio can be shown to equal one by plugging in the expression for the distribution of the item response data under the BCSM-MI given $\boldsymbol{\theta}_j$ (Equation (D3)) and given $\boldsymbol{\theta}_j^*$. It follows that

$$\begin{aligned}
R\left(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j\right) &= \frac{p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j^*, \boldsymbol{\Sigma}_j\right) \prod_k p\left(\mathbf{z}_{jk} \mid \boldsymbol{\theta}_j, a_{jk}, \boldsymbol{\Sigma}_{j_b}\right) p\left(a_{jk} \mid \boldsymbol{\theta}_j^*, \mathbf{z}_{jk}, \boldsymbol{\Sigma}_{j_b}\right)}{p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_j\right) \prod_k p\left(\mathbf{z}_{jk} \mid \boldsymbol{\theta}_j^*, a_{jk}, \boldsymbol{\Sigma}_{j_b}\right) p\left(a_{jk} \mid \boldsymbol{\theta}_j, \mathbf{z}_{jk}, \boldsymbol{\Sigma}_{j_b}\right)} \\
&= \frac{p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j^*, \boldsymbol{\Sigma}_j\right) p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j, \mathbf{a}_j, \boldsymbol{\Sigma}_{j_b}\right) p\left(\mathbf{a}_j \mid \boldsymbol{\theta}_j^*, \mathbf{z}_j, \boldsymbol{\Sigma}_{j_b}\right)}{p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_j\right) \prod_k p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j^*, \mathbf{a}_j, \boldsymbol{\Sigma}_{j_b}\right) p\left(\mathbf{a}_j \mid \boldsymbol{\theta}_j, \mathbf{z}_j, \boldsymbol{\Sigma}_{j_b}\right)} \\
&= \frac{\frac{p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j^*, \mathbf{a}_j, \boldsymbol{\Sigma}_{j_b}\right) p\left(\mathbf{a}_j \mid \sigma_a\right)}{p\left(\mathbf{a}_j \mid \boldsymbol{\theta}_j^*, \mathbf{z}_j, \boldsymbol{\Sigma}_{j_b}\right)} p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j, \mathbf{a}_j, \boldsymbol{\Sigma}_{j_b}\right) p\left(\mathbf{a}_j \mid \boldsymbol{\theta}_j^*, \mathbf{z}_j, \boldsymbol{\Sigma}_{j_b}\right)}{\frac{p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j, \mathbf{a}_j, \boldsymbol{\Sigma}_{j_b}\right) p\left(\mathbf{a}_j \mid \sigma_a\right)}{p\left(\mathbf{a}_j \mid \boldsymbol{\theta}_j, \mathbf{z}_j, \boldsymbol{\Sigma}_{j_b}\right)} p\left(\mathbf{z}_j \mid \boldsymbol{\theta}_j^*, \mathbf{a}_j, \boldsymbol{\Sigma}_{j_b}\right) p\left(\mathbf{a}_j \mid \boldsymbol{\theta}_j, \mathbf{z}_j, \boldsymbol{\Sigma}_{j_b}\right)} \\
&= \frac{p\left(\mathbf{a}_j \mid \sigma_a\right)}{p\left(\mathbf{a}_j \mid \sigma_a\right)} = 1.
\end{aligned} \tag{D4}$$

Cite this article: Fox, J.-P. (2026). Bayesian Covariance Modeling of Differential Item Functioning. *Psychometrika*, 1048–1068. https://doi.org/10.1017/psy.2026.10101

https://doi.org/10.1017/psy.2026.10101 Published online by Cambridge University Press